Policy Learning with Asymmetric Counterfactual Utilities*

Eli Ben-Michael† Kosuke Imai‡ Zhichao Jiang§

November 29, 2023

Abstract

Data-driven decision making plays an important role even in high stakes settings like medicine and public policy. Learning optimal policies from observed data requires a careful formulation of the utility function whose expected value is maximized across a population. Although researchers typically use utilities that depend on observed outcomes alone, in many settings the decision maker's utility function is more properly characterized by the joint set of potential outcomes under all actions. For example, the Hippocratic principle to "do no harm" implies that the cost of causing death to a patient who would otherwise survive without treatment is greater than the cost of forgoing life-saving treatment. We consider optimal policy learning with asymmetric counterfactual utility functions of this form that consider the joint set of potential outcomes. We show that asymmetric counterfactual utilities lead to an unidentifiable expected utility function, and so we first partially identify it. Drawing on statistical decision theory, we then derive minimax decision rules by minimizing the maximum expected utility loss relative to different alternative policies. We show that one can learn minimax loss decision rules from observed data by solving intermediate classification problems, and establish that the finite sample excess expected utility loss of this procedure is bounded by the regret of these intermediate classifiers. We apply this conceptual framework and methodology to the decision about whether or not to use right heart catheterization for patients with possible pulmonary hypertension.

Keywords: Hippocratic oath, minimax regret, partial identification, policy learning, principal stratification

*We acknowledge the partial support from Cisco Systems, Inc. (CG #2370386), National Science Foundation (SES–2051196), Sloan Foundation (Economics Program; 2020–13946), National Natural Science of China (Grant No. 12371285, 12292984), and Fundamental Research Funds for the Central Universities, Sun Yat-sen University (Grant No. 23hytd010). We also thank the IQSS's Alexander and Diviya Magaro Peer Pre-Review Program for feedback.

†Assistant Professor, Department of Statistics & Data Science and Heinz College of Information Systems & Public Policy, Carnegie Mellon University. 4800 Forbes Avenue, Hamburg Hall, Pittsburgh PA 15213. Email: ebenmichael@cmu.edu URL: ebenmichael.github.io

‡Professor, Department of Government and Department of Statistics, Harvard University. 1737 Cambridge Street, Institute for Quantitative Social Science, Cambridge MA 02138. Email: imai@harvard.edu URL: https://imai.fas.harvard.edu

§Professor, School of mathematics, Sun Yat-sen University, Guangzhou Guangdong 510275. Email: jiangzhch7@mail.sysu.edu.cn URL: https://zhichaoj-git.github.io

1 Introduction

The well-known Trolley Problem in ethics goes as follows:

> Edward is the driver of a trolley, whose brakes have just failed. On the track ahead of him are five people; the banks are so steep that they will not be able to get off the track in time. The track has a spur leading off to the right, and Edward can turn the trolley onto it. Unfortunately there is one person on the right-hand track. Edward can turn the trolley, killing the one; or he can refrain from turning the trolley, killing the five (Thomson, 1976, p. 206).

Should Edward turn the trolley? Is killing someone worse than letting them die? Such ethical dilemmas frequently confront us in moral and legal debates concerning various issues that range from abortion to self-driving cars (e.g., Foot, 1967; Lin, 2016). Similarly, in the ethics of modern medicine, the Hippocratic principle of "do no harm" remains influential (e.g., Jonsen, 1978; Smith, 2005; Wiens et al., 2019). In the language of utility theory, a physician may assign a utility loss of greater magnitude to the case where a new drug harms a patient than to the case where not providing the new drug leads to the failure to save a patient (e.g., Bordley, 2009).

These examples illustrate the potential applications of *asymmetric counterfactual* utilities that depend not only on the observed outcome, but also on the counterfactual outcome that could occur under a different action, and treat actions differently depending on their corresponding potential outcomes. Yet, to the best of our knowledge, the existing literature on data-driven decision making and algorithmic policy learning assumes that the decision maker's utility function only depends on the observed outcome.

In this paper, we develop the methodological framework for optimal policy learning with asymmetric counterfactual utilities, which includes standard utilities based on marginal outcomes as a special case. We show that in general, asymmetric counterfactual utilities lead to an unidentifiable expected utility function. Therefore, we partially identify the expected utility and propose to minimize the maximum expected utility loss relative to a particular comparison policy. We consider the maximum expected utility loss relative to constant policies such as always-treat and

never-treat policies as well as the oracle policy that has complete knowledge of the unidentifiable terms in the expected utility function. We demonstrate that one can learn minimax decision rules from observed data by solving intermediate classification problems. We also establish that the finite sample regret of this procedure is bounded by regret of these intermediate classifiers.

We use this framework to re-assess the use of Right Heart Catheterization (RHC), an invasive diagnostic tool (Connors et al., 1996). We learn decision rules based on clinical variables as we vary the asymmetry in the costs between failing to prevent a patient's death and causing it via RHC. These decision rules differ depending on whether we minimize the worst-case expected utility loss relative to a constant policy (always or never using RHC) or the oracle policy that uses RHC optimally. We inspect how the choice of utility function and comparator affect the learned decision rules, finding substantial variability based on these choices. Finally, we translate these findings into directly interpretable patient outcomes, exhibiting a trade-off between limiting the worst-case proportions of patients that the policy harms or fails to save.

Related Literature. Recent years have seen an increased interest in algorithmic policy learning from randomized control trials or observational data. Many of these approaches follow a similar structure. First, quantify the expected utility of a policy based on the marginal distributions of the potential outcomes. Then, show how to identify the expected utility or regret from observable data and find a policy that optimizes an empirical analog. These approaches typically use inverse propensity score weighting or double-robust methods for the identification and estimation steps (see Zhao et al., 2012; Kitagawa and Tetenov, 2018; Athey and Wager, 2021, among others). There is also a related literature that focuses on identifying and estimating optimal policies in settings with unmeasured confounding via instrumental variables (see Cui and Tchetgen Tchetgen, 2021; Qiu et al., 2021).

More immediately relevant to our discussion here, recent work builds off classical ideas in decision theory and treatment choice (e.g. Manski, 2004, 2005, 2011) and considers scenarios where we cannot point identify the expected utility function for possible policies. One strand of

work considers choosing between two treatments or fixed decision rules based on a finite sample of data, when treatment effects are partially identified (e.g. Stoye, 2012; Ishihara and Kitagawa, 2021; Yata, 2021). This work typically involves directly solving an empirical minimax regret problem, but does not consider optimization over classes of individualized policies.

In contrast, another line of work considers learning optimal individualized decision rules in situations where treatment effects are only partially identified, using an empirical risk minimization approach. These include settings with unmeasured confounding (e.g., Kallus and Zhou, 2021; Pu and Zhang, 2021; Han, 2021; Cui, 2021) or limited overlap between different treatment conditions (e.g., Ben-Michael et al., 2021; Zhang et al., 2023). D'Adamo (2023) considers a general setup where the conditional expected potential outcomes and treatment effects are partially identified. These approaches take a minimax approach at the population level, deriving the population-level minimum expected utility or maximum regret. They then treat the population-level maximum regret or negative minimum expected utility as a risk, and use empirical risk minimization approaches and propensity score weighting or double-robust methods as above. Our work is in the vein, estimating individualized treatment rules via empirical risk minimization. However, we consider a different setting where treatment effects *are* point identified, but the expected utility function is partially identified because it is a function of the proportion of units within each principal stratum—an unidentifiable quantity under standard designs.

Finally, Babii et al. (2021) also consider asymmetric utilities, but only using observed outcomes. In contrast, we consider asymmetric *counterfactual* utilities, which depend on potential outcomes and is a generalization of Babii et al.'s approach (see Appendix D for details).

Paper outline. The paper proceeds as follows. Section 2 describes the goal of policy learning with asymmetric counterfactual utilities and reviews the standard symmetric case. Section 3 discusses partial identification of the expected utility function and the minimax population policies relative to different alternatives. Section 4 then shows how to estimate such policies from data. Finally, Section 5 applies this framework to the use of RHC, and Section 6 concludes.

2 Preliminaries

In this section, we introduce the notation and assumptions used throughout this paper. We also discuss the nature of asymmetric counterfactual utilities before providing a brief review of policy learning with symmetric utilities, which is a special case of our proposed framework.

2.1 Notation and assumptions

Suppose that we have a simple random sample of n units from a super population $\mathcal{P}$ where each unit $i = 1, \ldots, n$ has a set of characteristics $X_i \in \mathcal{X}$. We consider a binary treatment assignment decision $D_i \in \{0, 1\}$, which can be made by either individual i or a policy maker. We assume that the outcome Y_i is binary with $Y_i = 1$ indicating a desirable outcome (e.g., survival) and $Y_i = 0$ representing an undesirable outcome (e.g., death). Under the assumption that there is only one version of treatment and no interference across units, we have two binary potential outcomes for each unit i where $Y_i(d) \in \{0, 1\}$ represents the potential outcome under the scenarios where the unit receives the decisions $D_i = d$ for $d = 0, 1$.

The setup implies that the observed outcome for unit i can be written as $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$ and the tuple of random variables $\{X_i, D_i, Y_i(1), Y_i(0)\}$ is assumed to be independently and identically distributed. Importantly, under this setting, each unit belongs to one of the four *principal strata* defined by the values of the two potential outcomes, i.e., $(Y_i(1), Y_i(0)) = (y_1, y_0) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ (Frangakis and Rubin, 2002). For example, the principal stratum $(y_1, y_0) = (1, 0)$ represents a group of units who would yield the desirable outcome only when they are treated, i.e., $D_i = 1$, whereas the principal stratum $(y_1, y_0) = (1, 1)$ indicates a group of units whose outcome is desirable regardless of the treatment decision. Since we never observe the two potential outcomes at the same time for any given unit, it is impossible to know which principal stratum each unit belongs to without additional assumptions.

Throughout this paper, for notational simplicity, we will drop the individual i subscript in

expressions involving expectations over the distribution. We will also assume the strong ignorability and strict overlap assumptions for observational studies and randomized control trials.

Assumption 1 (Strong ignorablity and strict overlap)**.** $\{Y(1), Y(0)\} \perp\!\!\!\perp D \mid X$ and there exists an $\eta > 0$ such that $\eta < d(x) < 1 - \eta$ for all $x \in \mathcal{X}$ where $d(x) \equiv \Pr(D = 1 \mid X = x)$ represents the propensity score.

The assumption allows us to identify the expected potential outcome under decision d given covariates x, denoted as $m(d, x) \equiv \mathbb{E}[Y(d) \mid X = x]$. However, it is impossible to identify the *principal score*, or the conditional probability of belonging to a principal stratum given covariates, defined as $e_{y_1 y_0}(x) \equiv \Pr(Y(1) = y_1, Y(0) = y_0 \mid X = x)$, because we do not observe the two potential outcomes at the same time for a given unit (Ding and Lu, 2017; Jiang et al., 2022).

2.2 Asymmetric counterfactual utilities

We focus on deterministic individualized policies $\pi : \mathcal{X} \to \{0, 1\}$ that assign a binary treatment decision to individual units according to their characteristics $X \in \mathcal{X}$. To learn optimal policies from the observed data, we consider a utility function $u(d; y_1, y_0)$ that encodes the utility for taking treatment decision d for a unit in principal stratum (y_1, y_0). Crucially, this utility function depends on the values of *both* potential outcomes. This contrasts with the standard utility function $u(d; y)$, which only depends on the realized potential outcome $Y_i(d) = y$ under the decision d. We measure the overall quality of a policy π by its expected utility (also called the *value* or *social welfare*),

$$
\begin{aligned}
V(\pi) &= \mathbb{E}\left[\sum_{y_1=0}^{1}\sum_{y_0=0}^{1} \mathbb{1}\{Y(1) = y_1, Y(0) = y_0\}\left\{u(0; y_1, y_0)(1 - \pi(X)) + u(1; y_1, y_0)\pi(X)\right\}\right] \\
&= \mathbb{E}\left[\sum_{y_1=0}^{1}\sum_{y_0=0}^{1} e_{y_1 y_0}(X)\pi(X)\left\{u(1; y_1, y_0) - u(0; y_1, y_0)\right\}\right] + \mathbb{E}\left[\sum_{y_1=0}^{1}\sum_{y_0=0}^{1} e_{y_1 y_0}(X) u(0; y_1, y_0)\right].
\end{aligned} \tag{1}
$$

This setup lets the utility vary across different counterfactual outcomes even when the treatment decision and the realized outcome are the same, allowing for a richer specification of the decision problem. For example, the disutility from assigning treatment to a patient that is harmed

by it (i.e., $Y(1) = 0$ and $Y(0) = 1$) can be larger than the disutility from assigning treatment to a patient for whom it is useless (i.e., $Y(1) = Y(0) = 0$), despite the fact that the realized outcome $Y(1)$ is identical in both cases. A standard utility function does not distinguish between these two cases, assigning each a value of $u(1, Y(1))$. This utility function also allows for asymmetry in the utility gain or loss from treating a unit across principal strata. Returning to the Hippocratic oath, we can choose the utility function such that the absolute magnitude of the utility loss for harming a patient through treatment ($|u(1;0,1) - u(0;0,1)|$) is greater than that of the utility gain when the same treatment benefits another patient ($|u(1;1,0) - u(0;1,0)|$).

To encode this, and focus on key ideas, we parameterize the utility function as follows:

(i) the utility gain associated with a "useful treatment," i.e., $(y_1, y_0) = (1, 0)$, is $u_g - c_g \equiv u(1;1,0) - u(0;1,0)$ (e.g., treating with a drug that would benefit the patient)

(ii) the utility loss associated with a "harmful treatment," i.e., $(y_1, y_0) = (0, 1)$, is $-u_l - c_l \equiv u(1;0,1) - u(0;0,1)$ (e.g., treating with a drug that would harm the patient)

(iii) the utility loss of treating with a "harmless treatment," i.e., $(y_1, y_0) = (1, 1)$, is $-c_1 = u(1;1,1) - u(0;1,1)$ (e.g., treating with a drug that would not harm the patient)

(iv) the utility loss of treating with a "useless treatment," i.e., $(y_1, y_0) = (0, 0)$, is $-c_0 = u(1;0,0) - u(0;0,0)$ (e.g., treating with a drug that would not benefit the patient)

The values c_g, c_l, c_1, c_0 denote the cost of administering the treatment $d = 1$ relative to not doing so $d = 0$ in each of the four principal strata. The values u_g and u_l represent the magnitude of the utility gain and loss for administering a useful and harmful treatment, respectively. In this setting, these utility values are known and fixed by the decision maker. Utility functions of this and more general forms have been considered in the literature on decision theory (see, e.g., Stefánsson, 2015; Bradley and Stefánsson, 2017). Our focus is, however, on the estimation of individualized decision rules under these asymmetric counterfactual utility functions.

	$Y_i(0) = 1$	$Y_i(0) = 0$
$Y_i(1) = 1$	Harmless $-c$	Useful $u_g - c$
$Y_i(1) = 0$	Harmful $-u_l - c$	Useless $-c$

Table 1: Asymmetric counterfactual utility gain/loss for treating each of the four principal strata, relative to not treating. Each cell corresponds to the principal stratum defined by the values of the two potential outcomes, $Y_i(1)$ and $Y_i(0)$. Each entry represents the utility gain/loss of treatment assignment, relative to no treatment, for a unit that belongs to the corresponding principal stratum $(u(1; y_1, y_0) - u(0; y_1, y_0))$. c is the cost of treatment assignment. We assume that $u_l, u_g > 0$ and $c \geq 0$. Symmetric utilities are a special case with $u_g = u_l$.

Throughout, we will assume that the costs are identical, i.e., $c_g = c_b = c_1 = c_0 = c$. In addition, we assume u_g and u_l are positive, and c is non-negative. Table 1 summarizes this asymmetric counterfactual utility structure. Fixing the costs to be identical amounts to restricting the utility loss from a harmless and useless treatment to be equal. Without this restriction there will be an additional asymmetry due to the different costs, which would not affect our development, except to make the notation more cumbersome and results less interpretable.

Note that our asymmetric counterfactual utilities include symmetric utilities based on observed outcomes as special cases, thereby generalizing the standard setting considered in the policy learning literature. In Appendix D, we further show that although it is possible to construct asymmetric utilities without using principal strata (Babii et al., 2021), doing so still implies some restrictions on the structure of the resulting counterfactual utilities and hence they are a special case of our framework.

We will primarily be comparing two policies rather than considering one in isolation. We begin by defining the *expected utility loss* of policy π relative to another policy ϖ as the difference in values, $V(\varpi) - V(\pi)$. Using the relations $m(1, x) = e_{11}(x) + e_{10}(x)$ and $m(0, x) = e_{11}(x) + e_{01}(x)$, we show in Appendix J that we can write the expected utility loss in a simplified form:

$$R_{e_{01}}(\pi, \varpi) \equiv V(\varpi) - V(\pi) = \mathbb{E}\left[(\varpi(X) - \pi(X))\{u_g \tau(X) + (u_g - u_l)e_{01}(X) - c\}\right]. \quad (2)$$

where $\tau(x) \equiv \mathbb{E}(Y(1) - Y(0) \mid X = x) = m(1, x) - m(0, x)$ is the conditional average treatment effect (CATE) given the covariates $X = x$. Comparing two policies to each other allows us to leave the baseline utility for not treating a unit in principal stratum (y_1, y_0), $u(0; y_1, y_0)$ unspecified. In Appendix B we directly consider the expected utility of a single policy, and connect choices of the baseline utility to our discussion below.

Three components contribute to the expected utility loss in Eqn (2). First, the difference in the expected treatment effects for those treated under policies ϖ and π, scaled by the utility gain for a useful treatment u_g, represents a symmetric component of the utility, where we compare the marginal benefits of the policies. The second component is an asymmetric adjustment term, and relates to the probability of belonging to the principal stratum for whom the treatment is harmful (i.e., $(y_1, y_0) = (0, 1)$). This can counteract the marginal benefit of treatment and is scaled by the difference between the utility gain for a useful treatment and the loss for a harmful treatment, i.e., $u_g - u_l$. The final term c corresponds to the difference in the overall costs of the two policies.

The first and third components, the difference in effects and costs, are point identifiable under Assumption 1. The second component, however, is only *partially identifiable* due to the unidentifiability of the principal score $e_{01}(\cdot)$. We use the e_{01} subscript for the expected utility loss in Eqn (2) to signify this fact. Therefore, we cannot pinpoint whether any policy is superior to any other policy in general. The remainder of this paper focuses on handling this ambiguity.

2.3 Policy learning with symmetric utilities: A review

Before discussing policy learning under asymmetric counterfactual utility functions, we briefly review policy learning with *symmetric* utilities — a special case of our framework — where the absolute magnitude of the utility gain when the treatment leads to a desirable outcome is equal to that of the expected utility loss when it leads to an undesirable outcome, i.e., $u_g = u_l$. In this case, a policy can make up for the loss from harming some units by the gain from benefiting other

units. This can be seen in the following simplified version of the expected utility loss in Eqn (2):

$$R_{\text{symm}}(\pi, \varpi) = \mathbb{E}\left[\{\varpi(X) - \pi(X)\}\{u_g\tau(X) - c\}\right]. \tag{3}$$

The symmetric utility does not involve the principal score $e_{10}(\cdot)$, and is identifiable under Assumption 1. Thus, under this setting, the oracle optimal policy that minimizes the expected utility loss relative to any other policy is $\pi^{\text{symm}}(x) \equiv \mathbb{1}\{u_g\tau(x) \geq c\}$. This oracle policy assigns the treatment to all individuals with characteristics x if their expected utility gain of assigning the treatment relative to not assigning it at least makes up for its cost, i.e., $u_g\tau(x) \geq c$. Note that this is equivalent to maximizing the value $V(\pi)$ directly.

Under Assumption 1, therefore, we can write the symmetric expected utility loss in Eqn (3) in terms of the observed data by using a scoring function $\Gamma_w(x, d, y)$ such that $\mathbb{E}\left[\Gamma_w(X, D, Y) \mid X = x\right] = m(w, x)$. For example, the Inverse Probability-of-treatment Weighting (IPW) scoring function uses the IP weighting function $\gamma_w(D, X) \equiv \frac{wD}{d(X)} + \frac{(1-w)(1-D)}{1-d(X)}$ to weight the observed outcome by the inverse probability of receiving the decision d: $\Gamma_w^{\text{ipw}}(X, D, Y) = Y\gamma_w(D, X)$. An alternative is the Doubly Robust (DR) scoring function that combines the observed outcomes and their conditional expectations: $\Gamma_w^{\text{dr}}(X, D, Y) = m(w, X) + \{Y - m(w, X)\}\gamma_w(D, X)$. With such a scoring rule, we can then write the symmetric expected utility loss function as: $\mathbb{E}\left[\{\varpi(X) - \pi(X)\}\left\{u_g(\Gamma_1(X, D, Y) - \Gamma_0(X, D, Y)) - c\right\}\right]$, where the observable quantity $\Gamma_1(X, D, Y) - \Gamma_0(X, D, Y)$ has replaced the causal quantity $\tau(X)$. See Knaus (2020) for a recent review.

In order to empirically find optimal policies from data, recent approaches estimate the propensity score $\hat{d}(\cdot)$ and/or the conditional expected potential outcome $\hat{m}(\cdot, \cdot)$ to create estimated scores $\widehat{\Gamma}(X_i, D_i, Y_i)$. For example, we can estimate the IP weights as $\hat{\gamma}_w(D, X) \equiv \frac{wD}{\hat{d}(X)} + \frac{(1-w)(1-D)}{1-\hat{d}(X)}$, the IPW scoring function as $\widehat{\Gamma}_w^{\text{ipw}}(X, D, Y) \equiv Y\hat{\gamma}_w(X, D)$, and the DR scoring function as $\widehat{\Gamma}_w^{\text{dr}}(X, D, Y) \equiv \hat{m}(w, X) + \{Y - \hat{m}(w, X)\}\hat{\gamma}_w(X, D)$. Then, we solve the sample analog of

Eqn (3). This leads to finding policy $\hat{\pi}$ that solves the following optimization problem:

$$\min_{\pi \in \Pi} \; -\frac{1}{n} \sum_{i=1}^{n} \pi(X_i) \left\{ u_g \left(\widehat{\Gamma}_1(X_i, D_i, Y_i) - \widehat{\Gamma}_0(X_i, D_i, Y_i) \right) - c \right\},$$

where Π represents the *policy class* and restricts the functional form of potential policies. Athey and Wager (2021) establish strong asymptotic guarantees on the regret of the empirical $\hat{\pi}$ relative to the best-in-class policy when using the DR approach with appropriately cross-fit models (see also Zhao et al., 2012; Kitagawa and Tetenov, 2018).

3 Policy learning with asymmetric counterfactual utilities

We now turn to the problem of finding optimal policies in the general asymmetric case where $u_g \neq u_l$. We will first consider the identification problems in the population — i.e., with infinite data. We then show how to learn policies empirically from observed data in Section 4. In Appendix A, we consider an alternative formulation as a constrained optimization problem.

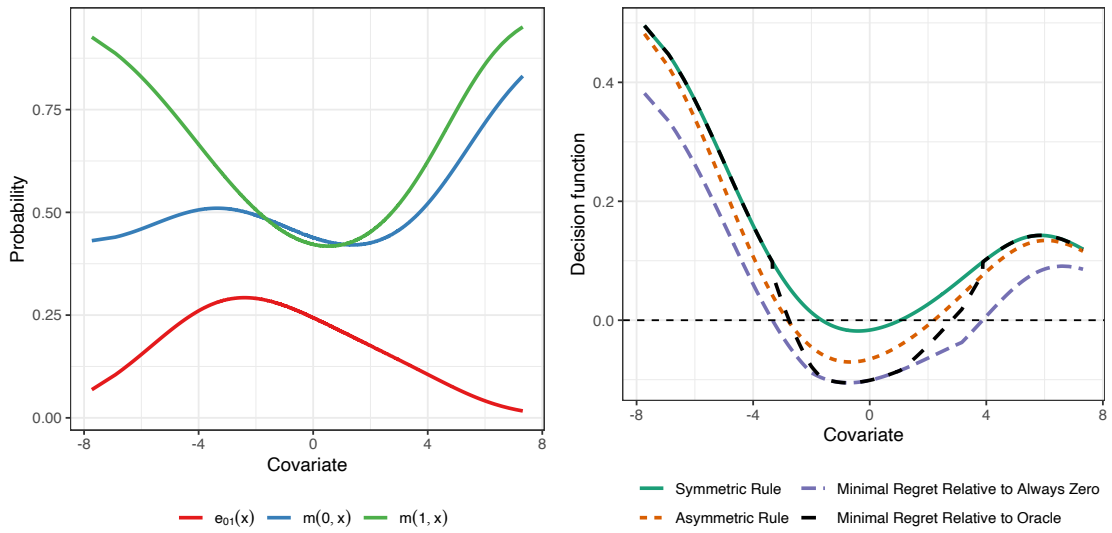
3.1 The oracle policy with an asymmetric counterfactual utility function

We begin by considering the oracle policy in the general asymmetric case. By direct computation, the (unconstrained) policy that has the maximal possible value with an asymmetric counterfactual utility function is given by:

$$\pi^o \equiv \operatorname*{argmax}_{\pi} V(\pi) = \mathbb{1} \left\{ \tau(\cdot) \geq \frac{u_l - u_g}{u_g} e_{01}(\cdot) + \frac{c}{u_g} \right\}. \tag{4}$$

We refer to this as the *oracle* policy, because it has access to the unknown (and generally unknowable) principal scores. Unlike in the symmetric case, this policy includes the principal score e_{01}, which is unidentifiable under Assumption 1. Since $0 \leq e_{01}(x) \leq 1$ for all x, the asymmetric oracle policy uses a varying threshold for assigning the treatment where the threshold depends on the principal score $e_{01}(X)$.

The way in which the oracle policy depends on the principal score is characterized in part



(a) Setting

(b) Decision rules

Figure 1: Example decision rules in a hypothetical example with a single covariate and cost $c = 0$. The left plot (a) presents the setting of the example where the values of the principal score $e_{01}(x) = \Pr(Y(1) = 0, Y(0) = 1 \mid X = x)$ are in red while the conditional expectations $m(d, x) = \mathbb{E}(Y(d) \mid X = x)$ are in blue for $d = 0$ and green for $d = 1$, respectively. The right plot (b) presents the decision rules corresponding to (i) the oracle in the symmetric case where $u_g = u_l$ (in green); (ii) the oracle in the asymmetric case where $u_g = 0.8$ and $u_l = 1$ (in orange); (iii) the minimal expected utility loss solution relative to always not assigning the treatment $d = 0$ (in purple); and (iv) the minimax regret solution relative to the oracle (in black). All rules have been transformed so that the policy takes decision 1 when the rule is greater than or equal to zero.

by the nature of the asymmetry in the utility function. Consider the case where treatment is costless ($c = 0$). When $u_l > u_g$, causing the undesirable outcome by assigning the treatment is considered worse than failing to prevent such an outcome by not providing the treatment. This raises the threshold for assigning the treatment because the expected effect must be larger in order to compensate for the downside risk of causing the undesirable outcome. As a result, this biases the oracle policy towards inaction. Conversely, when $u_l < u_g$, it is better to cause the undesirable outcome than fail to prevent it by inaction. In this case, the threshold for the treatment assignment is lower, biasing the oracle policy towards action.

Figure 1a shows a one-dimensional example of these decision rules where the cost is zero, $c = 0$. The principal score $e_{01}(x)$ is shown in red whereas the conditional expectations $m(d, x) = \mathbb{E}(Y(d) \mid X = x)$ are shown in blue ($d = 0$) and green ($d = 1$). Figure 1b shows the functions

that make up the decision rules in this example, centered so that the corresponding policies assign $d = 1$ if the function is positive. The symmetric case is shown in green, while the oracle in the asymmetric case ($u_g = 0.8$ and $u_l = 1$) is in orange. This plot also shows two other solutions discussed in Section 3.3; the minimal expected utility loss solution relative to always not assigning the treatment (purple), and the minimal regret solution relative to the oracle (black).

In this example, providing the treatment $d = 1$ leads to a higher probability of the desirable outcome except near zero. Therefore, with a symmetric utility function, the oracle policy would assign the treatment in most cases (green). However, the asymmetric case is different (orange). There is a region of the covariate space where the principal score $e_{01}(x)$ is relatively high, leading to a sufficiently high probability that the treatment causes the undesirable outcome. Therefore, the asymmetric oracle rule has a higher threshold for the treatment assignment, only providing the treatment when the CATE $\tau(x)$ is large enough and the principal score $e_{10}(x)$ is small enough.

3.2 Partial identification and minimizing worst-case expected utility loss

Recall that the unidentifiability of the principal score $e_{01}(x)$ for any $x \in \mathcal{X}$ makes it impossible to identify the expected utility loss in Eqn (2) in the general asymmetric case with $u_g \neq u_l$. However, we can *partially identify* the principal score by deriving its sharp upper and lower bounds, $L(x)$ and $U(x)$. We then take a minimax approach, and find the policy π^* in the policy class Π that minimizes the maximal expected utility loss relative to an alternative policy ϖ:

$$\pi^* \in \arg\min_{\pi \in \Pi} R_{\text{sup}}(\pi, \varpi) \quad \text{where} \quad R_{\text{sup}}(\pi, \varpi) = \max_{e_{01}(x) \in [L(x), U(x)]} R_{e_{01}}(\pi, \varpi). \tag{5}$$

Note that the maximum expected utility loss $R_{\text{sup}}(\pi, \varpi)$ is relative to a *particular* alternative policy ϖ, and is maximal over all possible values for the principal score $e_{01}(x)$. As we show below, the choice of this alternative policy will lead to different objectives and optimal solutions (see Cui, 2021, for a recent general discussion).

Eqn (5) is an example of a treatment choice problem under ambiguity (Manski, 2005, 2011).

Such minimax formulations of the problem have been widely considered in settings where value functions depend on the marginal distribution of the potential outcomes, but the CATE $\tau(x)$ is not point identified (e.g. Manski, 2007; Stoye, 2012; Kallus, 2018; Ben-Michael et al., 2021; Ishihara and Kitagawa, 2021; Yata, 2021; Zhang et al., 2023; D'Adamo, 2023). A key distinction between Eqn (5) and these other problems, however, is that in our setting the value function depends on the principal score $e_{10}(x)$, which is not point identified even in randomized control trials.

To derive the sharp lower and upper bounds of the principal score, we first define two classification functions:

$$\delta_+(x) \;=\; \mathbb{1}\{m(0,x) + m(1,x) - 1 \geq 0\} \quad \text{and} \quad \delta_\tau(x) \;=\; \mathbb{1}\{\tau(x) \geq 0\}. \tag{6}$$

Notice that the difference in the probability that both potential outcomes are one and zero is given by $e_{11}(x) - e_{00}(x) = m(0,x) + m(1,x) - 1$, which is the decision function for classifier $\delta_+(x)$. In other words, we have $\delta_+(x) = 1$ if and only if $e_{00}(x) \leq e_{11}(x)$. Thus, we can view $\delta_+(x)$ as classifying whether there is a higher probability that both potential outcomes are one rather than zero. In contrast, noting that $\tau(x) = e_{10}(x) - e_{01}(x)$, $\delta_\tau(x)$ classifies whether there is a higher probability that the treatment is useful rather than harmful. This corresponds to the symmetric oracle rule with cost $c = 0$.

With these classifiers, we can use the Fréchet bounds to find the sharp lower and upper bounds for the principal score, $e_{01}(x) \in [L(x), U(x)]$ for all x (e.g. Heckman et al., 1997; Jiang et al., 2016; Kallus, 2018):

$$L(x) \;=\; \max\{0, 1 - m(1,x) + m(0,x) - 1\} \;=\; \max\{0, -\tau(x)\} \;=\; -\tau(x)\{1 - \delta_\tau(x)\} \tag{7}$$

$$U(x) \;=\; \min\{m(0,x), 1 - m(1,x)\} \;=\; m(0,x) + \delta_+(x)\{1 - m(0,x) - m(1,x)\}. \tag{8}$$

These lower and upper bounds are sharp (Rüschendorf, 1981) and are point-identifiable from observable data. With them we can create a point-identifiable objective.

3.3 Worst case expected utility loss relative to different alternative policies

We now inspect the worst-case expected utility loss $R_{\text{sup}}(\pi, \varpi)$ in Eqn (5) for different choices of alternative policy ϖ. We consider three main alternatives. First, the "never-treat" policy $\pi^{\mathbb{O}}$ that does not treat anyone, i.e., $\pi^{\mathbb{O}}(x) = 0$ for all x. Second, the "always-treat" policy $\pi^{\mathbb{1}}$ that treats everyone, i.e., $\pi^{\mathbb{1}}(x) = 1$ for all x. In many cases, the alternative to algorithmic decision making via a data-driven policy is to take the same decision for everyone; thus these two policies are of interest as they represent the standard of care in the absence of an individualized policy (see Appendix B for connections to maximin policies that maximize the minimum expected utility). We will denote the policies that minimize these worst-case losses as $\pi^*_{\mathbb{O}} \equiv \arg\min_\pi R_{\text{sup}}(\pi, \pi^{\mathbb{O}})$ and $\pi^*_{\mathbb{1}} \equiv \arg\min_\pi R_{\text{sup}}(\pi, \pi^{\mathbb{1}})$.

Finally, we consider the *minimax regret policy* that minimizes the worst case expected utility relative to the value of the best-possible policy that has access to the principal scores $e_{01}(\cdot)$. Formally, the minimax regret policy is defined as $\pi^*_o \equiv \arg\min_\pi \max_{e_{01}(x)\in[L(x),U(x)]} \max_{\pi'} R_{e_{01}}(\pi, \pi')$. The definition of the oracle policy π^o above implies that this is equivalent to choosing π^o as the alternative policy. That is, $\pi^*_o = \arg\min_\pi R_{\text{sup}}(\pi, \pi^o)$ where $R_{\text{sup}}(\pi, \pi^o) = \max_{e_{01}(x)\in[L(x),U(x)]} \max_{\pi'} R_{e_{01}}(\pi, \pi')$ is the *regret* of policy π.

Minimax regret policies are often studied in the policy learning literature because alternatives, such as maximin policies, tend to be too conservative (see e.g., Manski, 2007, 2011; Stoye, 2012; Yata, 2021, among many others). Note that when defining the minimax regret policy across a constrained policy class, we compare to the best possible *unconstrained* policy, i.e., $\arg\min_{\pi\in\Pi} \max_{e_{01}(x)\in[L(x),U(x)]} \max_{\pi'} R_{e_{01}}(\pi, \pi') = \arg\min_{\pi\in\Pi} R_{\text{sup}}(\pi, \pi^o)$. The resulting policy will be different in general from the policy that minimizes the regret relative to the best-in class policy, and the unconstrained form of the regret will be larger.

The following theorem shows that the worst-case expected utility loss relative to each of these three policies takes a common form.

Theorem 3.1 (Worst case expected utility loss)**.** Let $\pi : \mathcal{X} \to \{0, 1\}$ be a deterministic policy. For comparison policy $\varpi \in \{\pi^{\mathbb{O}}, \pi^{\mathbb{1}}, \pi^{o}\}$, the worst-case expected utility loss of π relative to ϖ is

$$\begin{aligned} R_{\sup}(\pi, \varpi) &= C - \mathbb{E}\left[\pi(X)\left\{c_1^{\varpi}(X)m(1, X) + c_0^{\varpi}(X)m(0, X) + c^{\varpi}(X)\right\}\right] \\ &= C - \mathbb{E}\left[\pi(X)\left\{c_1^{\varpi}(X)\Gamma_1(X, D, Y) + c_0^{\varpi}(X)\Gamma_0(X, D, Y) + c^{\varpi}(X)\right\}\right], \end{aligned} \tag{9}$$

where C is a constant that does not depend on π, and $c_1^{\varpi}(\cdot), c_0^{\varpi}(\cdot), c^{\varpi}(\cdot) : \mathcal{X} \to \mathbb{R}$ are functions that depend on $\delta_+(\cdot), \delta_\tau(\cdot), \pi^*_{\mathbb{O}}$, or $\pi^*_{\mathbb{1}}$.

The maximum expected utility loss objective in Theorem 3.1 is a weighted average of the expected potential outcomes under treatment and no treatment plus a proxy for the cost. The choice of alternative policy ϖ determines these weights $c_0^{\varpi}(\cdot), c_1^{\varpi}(\cdot)$ and cost $c^{\varpi}(\cdot)$, all of which potentially vary with the covariates X; we give explicit formulas for these functions in Appendix H. Note that the special case of a symmetric utility (Section 2.3) is also of this form, with $c_1^{\varpi}(X) = -c_0^{\varpi}(X) = u_g$ and $c^{\varpi}(X) = c$. Similarly, the two classifiers in Eqn (6) have this form, with δ_τ corresponding to $c_1^{\varpi}(X) = -c_0^{\varpi}(X) = 1$ and $c^{\varpi}(X) = 0$, and δ_+ corresponding to $c_1^{\varpi}(X) = c_0^{\varpi}(X) = 1$ and $c^{\varpi}(X) = -1$.

The second line of Eqn (H.1) shows how to write the worst-case expected utility loss $R_{\sup}(\pi, \varpi)$ in terms of observable data using the scoring functions Γ_w (either IPW or DR) discussed in Section 2.3. So, targeting the worst-case expected utility loss yields an objective function that is identifiable, unlike the true expected utility loss. As shown below, this allows us to construct decision rules based on observable data that control the true expected utility loss by minimizing the worst-case expected utility loss.

Constructing a utility function based on principal strata allows decision makers to define their goals directly in terms of individualized notions of useful and harmful treatments. Nevertheless, Theorem 3.1 shows that the minimax expected utility loss problem reduces to a decision problem that only involves the marginal distribution of the potential outcomes. The principal score $e_{01}(x)$ will not be involved in the remaining estimation strategies and results, having been replaced with point-identifiable upper and lower bounds.

However, the weighting and cost functions induced by the utility function and choice of alternative policy correspond to a covariate-dependent asymmetry in terms of the *marginal* potential outcomes. Depending on the values of the nuisance classifiers, Eqn (H.1) places more or less weight on outcomes under treatment versus outcomes under control. Thus, Eqn (H.1) is related to the covariate-dependent loss minimization problem considered by Babii et al. (2021) that depends on marginal outcomes, even though it was derived from placing an asymmetric counterfactual utility on the principal strata. A key distinction is that because Eqn (H.1) involves the unknown nuisance classifiers, we must estimate the corresponding loss function. We analyze the consequences of this in Section 4.2.

Finally, note that here we restrict to *deterministic* policies to derive the form of the minimax expected utility loss in Theorem 3.1. As Cui (2021) discusses, unlike with the expected utility loss relative to the always treat or never treat policies, allowing for *stochastic* policies that randomize between actions can lead to lower loss, though this leads to a more complicated form. We leave further understanding the implications for stochastic policies to future work.

Next, we compute and inspect the policy that is the unconstrained minimizer of the maximum expected utility loss in the population, relative to each of the three alternative policies in turn. We will then turn to estimating constrained policies in finite samples in Section 4 below.

3.3.1 Expected utility loss relative to a constant decision

We begin by considering the worst-case expected utility loss relative to the never-treat policy.

Corollary 3.2 (Minimax expected utility loss relative to the never-treat policy)**.** If $u_g \geq u_l$, the solution to Eqn (5), $\pi^*_{\mathbb{O}} \equiv \arg\min_\pi R_{\text{sup}}(\pi, \pi^{\mathbb{O}})$ is the symmetric policy,

$$\pi^*_{\mathbb{O}}(x) = \mathbb{1}\left\{\tau(x) \geq \frac{c}{u_g}\right\} = \pi^{\text{symm}}(x).$$

Otherwise, if $u_g < u_l$, it is given by,

$$\pi^*_{\mathbb{O}}(x) = \begin{cases} \mathbb{1}\left\{m(1,x) \geq \frac{u_l}{u_g} m(0,x) + \frac{c}{u_g}\right\}, & \delta_+(x) = 0, \\ \mathbb{1}\left\{m(1,x) \geq \frac{u_g}{u_l} m(0,x) + \frac{u_l - u_g + c}{u_l}\right\}, & \delta_+(x) = 1. \end{cases}$$

Corollary 3.2 shows that the form of the minimax expected utility loss policy depends on the direction of the asymmetry. To build intuition, consider the case where the treatment is costless ($c = 0$). If $u_g > u_l$ — so we would rather cause an undesirable outcome than to fail to prevent it — then the minimax solution relative to the never-treat policy is the same as the optimal rule under a symmetric utility function: assign the treatment when the CATE is positive. In this case, the unit is more likely to be in the $(y_1, y_0) = (1, 0)$ stratum than the $(y_1, y_0) = (0, 1)$ stratum, and since $u_g > u_l$, it will be better to treat the unit than to not. Conversely, when the CATE is negative it may still be better to treat the unit, but in the worst case it is not. To minimize the worst-case expected utility loss relative to never treating, the minimax loss policy does not treat.

However, the minimax solution is different when $u_g < u_l$ — i.e., when it is worse to cause an undesirable outcome than to fail to prevent it. In this case, the oracle rule depends on the value of the classifier $\delta_+(x) = \mathbb{1}\{e_{00}(x) \leq e_{11}(x)\}$. If both potential outcomes are more likely to be zero than one, then the policy only treats if the probability that $Y(1)$ equals one is higher than the probability that $Y(0)$ equals one by a factor of $\frac{u_l}{u_g} > 1$. Comparing to the decision rule under the symmetric utility, we see that this raises the threshold for assigning the treatment.

In contrast, if both potential outcomes are more likely to be one than zero, the threshold is raised by adding a constant cost $\frac{u_l - u_g}{u_l} > 0$, but the multiplicative factor on the probability that $Y(0)$ equals one is $\frac{u_g}{u_l} < 1$. Overall, this has the effect of creating a more cautious policy that provides the treatment less often.

Figure 1b shows the minimax decision rule relative to $\pi^{\mathbb{O}}$ (purple) in the one-dimensional example where $u_g < u_l$ and $c = 0$ — i.e., it is worse to cause an undesirable outcome than fail to prevent it. In this case, we see that the decision function is well below the symmetric rule shown in green (i.e. the CATE), leading to a large part of the covariate space being assigned no treatment even though the CATE is positive. In fact, this policy is overly cautious: it does not assign the treatment even in many cases where the oracle rule that knows the principal score would provide the treatment. This is because the alternative policy is to never treat anyone.

Appendix H shows the result for the minimax expected utility loss policy relative to the always-treat policy, which is more aggressive than the symmetric policy. It is the mirror image of the minimax loss policy relative to the never-treat policy, with the relation to u_g and u_l reversed.

3.3.2 The minimax regret policy

We next consider the policy that minimizes the expected utility loss relative to the oracle π^o in Eqn (4), or, equivalently, that minimizes the regret. For simplicity, we assume zero cost, i.e., $c = 0$; when $c > 0$, there will be further terms (see the proof of Theorem 3.1 in Appendix J).

Corollary 3.3 (Minimax regret policy)**.** When $c = 0$, the minimax regret policy for $u_g \geq u_l$ is given by,

$$\pi_o^*(x) = \begin{cases} 1, & \delta_\tau(x) = 1, \\ 0, & \pi_{\mathbb{1}}^*(x) = 0, \\ \mathbb{1}\left\{m(1,x) \geq \frac{2u_l}{u_g+u_l} m(0,x)\right\}, & \delta_\tau(x) = 0, \delta_+(x) = 0, \\ \mathbb{1}\left\{m(1,x) \geq \frac{u_g+u_l}{2u_l} m(0,x) + \frac{u_l-u_g}{2u_l}\right\}, & \delta_\tau(x) = 0, \delta_+(x) = 1, \end{cases}$$

and for $u_g < u_l$ the minimax regret policy is given by,

$$\pi_o^*(x) = \begin{cases} 1, & \pi_{\mathbb{O}}^*(x) = 1, \\ 0, & \delta_\tau(x) = 0, \\ \mathbb{1}\left\{m(1,x) \geq \frac{u_g+u_l}{2u_g} m(0,x)\right\}, & \delta_\tau(x) = 1, \delta_+(x) = 0, \\ \mathbb{1}\left\{m(1,x) \geq \frac{2u_g}{u_g+u_l} m(0,x) + \frac{u_l-u_g}{u_g+u_l}\right\}, & \delta_\tau(x) = 1, \delta_+(x) = 1. \end{cases}$$

Corollary 3.3 shows that we can write the worst-case regret and the minimax regret policy in terms of observable data, just as for the constant policies above. But, doing so requires *four* classifiers rather than one: (i) δ_+, which classifies whether $e_{00}(x) \leq e_{11}(x)$; (ii) δ_τ, which classifies whether the CATE is positive; (iii) the minimax loss solution relative to $\pi^{\mathbb{1}}$ and (iv) the minimax loss solution relative to $\pi^{\mathbb{O}}$. Recall from Corollary 3.2 that either $\pi_{\mathbb{O}}^*$ (when $u_g \geq u_l$) or $\pi_{\mathbb{1}}^*$ (when $u_g < u_l$) is the symmetric policy π^{symm}. Therefore, if the cost $c = 0$ as in Corollary 3.3, we only need *three* classifiers to construct the objective, since $\pi^{\text{symm}} = \delta_\tau$ in this case.

Inspecting the minimax solution relative to the oracle policy when $u_g \geq u_l$, we see that it assigns the treatment if the symmetric rule does, whereas it does not provide the treatment if the

minimax solution relative to the always-treat policy does not. In between these two extremes, the decision rule lowers the threshold for the treatment assignment relative to the symmetric rule. The opposite is true when $u_g < u_l$. If the symmetric rule does not assign treatment, the minimax solution relative to the oracle does not either, but it does provide the treatment whenever the minimax solution relative to the never-treat policy does. In between these two cases, the threshold for treatment assignment is higher than that under the symmetric rule.

Figure 1b shows the decision rule (black) in our running one-dimensional example where $u_g < u_l$. The decision rule is equivalent to $\pi^*_{\mathbb{O}}$ (purple) when the CATE is negative, and is equal to the CATE decision rule δ_τ when $\pi^*_{\mathbb{O}}(x) = 1$. When there is disagreement between the CATE rule δ_τ and $\pi^*_{\mathbb{O}}$, the minimax oracle rule interpolates between them, leading to a more aggressive policy that treats more individuals than $\pi^*_{\mathbb{O}}$. Comparing to the oracle rule (orange), we see that this interpolation causes the decision thresholds for the minimax oracle rule to be close to the best possible decision thresholds.

4 Learning a policy from data

Having established the behavior and form of the minimax loss policy π^* in Eqn (5) in the population for an unconstrained policy class, we now turn to the problem of learning a policy $\hat{\pi}$ from observed data within a constrained policy class Π.

4.1 Estimation algorithms

To begin, note that in finite samples we know neither the true scoring functions Γ_w nor the true weighting and cost functions $c_1^{\varpi}(\cdot), c_0^{\varpi}(\cdot), c^{\varpi}(\cdot)$—which depend on the nuisance classifiers—and so they must be estimated from data. As mentioned in Section 2.3, we can obtain estimates of the DR score $\widehat{\Gamma}_w^{\text{dr}}$ by plugging in estimates of the nuisance components. Similarly, with estimates of the nuisance classifiers, we can directly obtain estimates of the weighting and cost functions $\hat{c}_1^{\varpi}(\cdot), \hat{c}_0^{\varpi}(\cdot)$, and $\hat{c}^{\varpi}(\cdot)$ by plugging in to the formulas in Theorem 3.1.

This leads to the following procedure. First, obtain estimates of the nuisance components $\hat{m}$ and $\hat{d}$ and construct the DR scores. Then, estimate the nuisance classifiers and follow Theorem 3.1 to construct estimates of the weighting and cost functions. To find a policy relative to either the always-treat (if $u_g \geq u_l$) or never-treat (if $u_g < u_l$) policies, we estimate a single nuisance classifier, $\hat{\delta}_+$. Finding a policy relative to the oracle involves estimating *three* or *four* nuisance classifiers: $\hat{\delta}_+$, $\hat{\delta}_\tau$, and the minimax loss policies relative to never and always treating, $\hat{\pi}_{\mathbb{O}}$ and $\hat{\pi}_{\mathbb{1}}$. With these in hand, we then find a data-driven policy $\hat{\pi}$ that solves the following optimization problem (dropping the constant that does not depend on the policy π):

$$\hat{\pi} \in \underset{\pi \in \Pi}{\arg\min}\ \hat{R}_{\text{sup}}(\pi, \varpi) \tag{10}$$

where $\hat{R}_{\text{sup}}(\pi, \varpi) = -\frac{1}{n}\sum_{i=1}^{n} \pi(X_i)\left\{\hat{c}_1^{\varpi}(X_i)\widehat{\Gamma}_1^{\text{dr}}(X_i, D_i, Y_i) + \hat{c}_0^{\varpi}(X_i)\widehat{\Gamma}_0^{\text{dr}}(X_i, D_i, Y_i) + \hat{c}^{\varpi}(X_i)\right\}$.

There are two ways to estimate the nuisance classifiers. The first is an *empirical risk minimzation* approach, where we solve Eqn (10) with the appropriate weighting and cost functions.[1] Appendix C explicitly details this procedure. As shown in Section 4.2 below, the estimated nuisance classifiers must have low regrets relative to the true ones in order for our learned policy $\hat{\pi}$ to have low worst-case expected utility loss; therefore, we must choose a flexible policy class. This is in contrast to estimating our policy of interest $\hat{\pi}$, whose performance we measure relative to the best possible constrained policy. An alternative is to take a *plug-in* approach, using our estimates of the conditional expectation function $\hat{m}$ to directly create estimates of the classifier; e.g., $\hat{\delta}_+(x) = \mathbb{1}\{\hat{m}(1,x) + \hat{m}(0,x) \geq 1\}$ and $\hat{\delta}_\tau(x) = \mathbb{1}\{\hat{m}(1,x) - \hat{m}(0,x) \geq 0\}$.

4.2 Excess worst-case expected utility loss

To understand the statistical properties of our learned minimax policy $\hat{\pi}$, we will compare it to the policy π^* that minimizes the worst-case expected utility loss in the population among those in the policy class Π by solving Eqn (5). For a given alternative policy ϖ, we will use the excess worst-

[1] For the nuisance classifiers δ_+ and δ_τ, the weighting and cost functions are known, and so need not be estimated.

case expected utility loss $R_{\text{sup}}(\hat{\pi}, \varpi) - R_{\text{sup}}(\pi^*, \varpi)$ to measure the quality of the learned minimax loss policy $\hat{\pi}$ since $R_{\text{sup}}(\pi^*, \varpi)$ is the best possible expected utility loss in the worst case. We assume that the nuisance components and classifiers have been obtained from a separate sample, and so can be treated as fixed for our finite sample results. However, our results can be extended to solving Eqn (10) by cross-fitting nuisance components and classifiers to obtain the estimates of $\widehat{\Gamma}^{\text{dr}}_w(X_i, D_i, Y_i)$ and $\hat{c}^{\omega}_w(X_i)$ (see Athey and Wager, 2021, and Appendix F.1).

To state our results, we define several new quantities. First, we measure the quality of the estimated nuisance classifiers, $\hat{\delta}_+$ and $\hat{\delta}_\tau$, by their regrets,

$$R_+(\hat{\delta}_+) \equiv \mathbb{E}\left[\mathbb{1}\{\hat{\delta}_+(X) \neq \delta_+(X)\}\,|m(1,X) + m(0,X) - 1|\right]$$
$$R_\tau(\hat{\delta}_\tau) \equiv \mathbb{E}\left[\mathbb{1}\{\hat{\delta}_\tau(X) \neq \delta_\tau(X)\}\,|m(1,X) - m(0,X)|\right],$$

where $\hat{\delta}_+$ and $\hat{\delta}_\tau$ are treated as fixed and the covariate X is random. Second, we measure the complexity of the policy class Π by its ability to overfit to noise via the *population Rademacher complexity*

$$\mathcal{R}_n(\Pi) \;\equiv\; \mathbb{E}_{X,\varepsilon}\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\pi(X_i)\right|\right],$$

where ε_i are i.i.d. random variables with $\Pr(\varepsilon_i = 1) = \Pr(\varepsilon_i = -1) = 1/2$, and the expectation is taken over both ε_i and X_i (Wainwright, 2019, §4).

We now present two finite sample bounds on the excess worst case expected utility loss, one for learning a minimax loss policy relative to the always or never treat policies (Theorem 4.1), and the other for learning a minimax loss policy relative to the oracle (Theorem 4.2).

Theorem 4.1. Let $\hat{\pi}$ solve Eqn (10) with alternative policy $\varpi = \pi^{\mathbb{O}}$ (if $u_g < u_l$) or $\varpi = \pi^{\mathbb{1}}$ (if $u_g \geq u_l$), and with nuisance functions $\hat{m}$ and $\hat{d}$ and classifier $\hat{\delta}_+$ fit on a separate sample. Let π^* solve the population problem in Eqn (5). The excess worst-case expected utility loss of $\hat{\pi}$ relative to π^* satisfies

$$R_{\text{sup}}(\hat{\pi}, \varpi) - R_{\text{sup}}(\pi^*, \varpi) \leq 2U \times \left\{\frac{6+\eta}{\eta} \times \left(2\mathcal{R}_n(\Pi) + \frac{t}{\sqrt{n}}\right) + \sum_{w=0}^{1} \|\hat{\gamma}_w - \gamma_w\|_2 \, \|\hat{m}(w,\cdot) - m(w,\cdot)\|_2\right\}$$
$$+ (u_g - u_l) \times \left\{R_+(\hat{\delta}_+) + \frac{t}{2\sqrt{n}}\right\},$$

with probability at least $1 - 2\exp\left(-\frac{t^2}{2}\right)$, where η is the overlap parameter in Assumption 1, U is a constant depending on the utility values, $\|\hat{\gamma}_w - \gamma_w\|_2^2 \equiv \mathbb{E}\left[\{\hat{\gamma}_w(D,X) - \gamma_w(D,X)\}^2\right]$ and $\|\hat{m}(w,\cdot) - m(w,\cdot)\|_2^2 \equiv \mathbb{E}[\{\hat{m}(w,X) - m(w,X)\}^2]$.

Theorem 4.2. Let $\hat{\pi}_o$ solve Eqn (10) with alternative policy set to be the oracle, $\varpi = \pi^o$, and with nuisance functions $\hat{m}$ and $\hat{d}$ and classifiers $\hat{\delta}_+, \hat{\delta}_\tau, \hat{\pi}_{\mathbb{O}}$, and $\hat{\pi}_{\mathbb{1}}$ fit on a separate sample. Let π_o^* solve the population problem in Eqn (5). The excess worst-case regret of $\hat{\pi}_o$ relative to π_o^* satisfies

$$
\begin{aligned}
R_{\text{sup}}(\hat{\pi}_0, \pi^o) - R_{\text{sup}}(\pi_o^*, \pi^o) \leq U \times & \left(\frac{6+\eta}{\eta} \times \left(2\mathcal{R}_n(\Pi) + \frac{t}{\sqrt{n}} \right) + \|\hat{\gamma} - \gamma\|_2 \sum_{w=0}^{1} \|\hat{m}(w,\cdot) - m(w,\cdot)\|_2 \right) \\
& + 2 \times \left(R_{\text{sup}}(\hat{\pi}_{\mathbb{1}}, \pi^{\mathbb{1}}) - R_{\text{sup}}(\pi_{\mathbb{1}}^*, \pi^{\mathbb{1}})\right) + 2 \times \left(R_{\text{sup}}(\hat{\pi}_{\mathbb{O}}, \pi^{\mathbb{O}}) - R_{\text{sup}}(\pi_{\mathbb{O}}^*, \pi^{\mathbb{O}})\right) \\
& + (u_g - u_l) \times \left(R_+(\hat{\delta}_+) + R_\tau(\hat{\delta}_\tau) + \frac{t}{2\sqrt{n}} \right),
\end{aligned}
$$

with probability at least $1 - 2\exp\left(-\frac{t^2}{2}\right)$, where U is a constant depending on the utility values.

Theorems 4.1 and 4.2 reveal three reasons why the data-specific policy $\hat{\pi}$ can differ from the population policy π^*. First, as captured via the Rademacher complexity term, even if the outcome model, propensity score model, and nuisance classifiers were all known, $\hat{\pi}$ could simply over fit to noisy data. Fortunately, we can choose the complexity of Π and often prefer a relatively simple policy class for its interpretability and transparency. The results above will be relative to the best possible policy in the selected policy class. Thus, we could control this by limiting the complexity of our search space. For example, if the policy class Π has a finite VC dimension ν, the Rademacher complexity scales like $\mathcal{R}_n(\Pi) = O\left(\sqrt{\frac{\nu}{n}}\right)$ (Wainwright, 2019, §5).

Second, there is error in our estimates of the outcome and propensity score models. However, following Athey and Wager (2021), using the DR scores protects against this error; only the product of the errors enter the bound, which decreases faster than $1/\sqrt{n}$ under typical assumptions. These two sources of error occur in symmetric policy learning problems. In the symmetric case when $u_g = u_l$ (and so $\pi^{\mathbb{1}} = \pi^{\mathbb{O}} = \pi^o$), Theorem 4.1 is a special case of the results in Athey and Wager (2021).

Finally, there is error in the nuisance classifiers, which is particular to our setting.[2] For the

[2]This type of error structure appears in other policy learning settings with partial identification (D'Adamo, 2023).

minimax loss policy relative to never or always treating, this error is measured by the regret for $\hat{\delta}_+$: if it correctly classifies cases that are not very close to the decision boundary (i.e. $|m(0,x)+m(1,x)-1|$ is not near zero), this component will be small. Similarly for the minimax loss policy relative to the oracle, there are additional terms from the regret of $\hat{\delta}_\tau$ and the excess worst case expected utility for the minimax loss policies relative to always and never treating.

If we estimate the nuisance classifiers via empirical risk minimization, results from Kitagawa and Tetenov (2018); Athey and Wager (2021) (and Theorem 4.1 for $\hat{\pi}_{\mathbb{1}}$ and $\hat{\pi}_{\mathbb{0}}$) imply that the regret will primarily be controlled by the complexity of the policy classes we optimize over for the nuisance classifiers. Unlike for the minimax loss policy class Π, unless the nuisance classifier class contains the *true* function, there will be irreducible approximation error in the misclassification term. Therefore, we might choose more complex classes, in which case the regret of the nuisance classifiers will primarily control the overall excess expected utility loss.

To analyze the plug-in approach, we use a different characterization of the complexity of the learning problem: the proportion of cases that are close to the decision boundary. Focusing on the nuisance classifier δ_+, we follow Audibert and Tsybakov (2007) and characterize this via the following *margin condition*.

Assumption 2 (Margin condition)**.** There exists an $\alpha > 0$ and a constant C such that for any $t \geq 0$, $\Pr(|m(1,X)+m(0,X)-1| \leq t) \leq Ct^\alpha$.

The margin parameter α determines how many cases are allowed to be close to the boundary, with a larger value leading to a stronger assumption that fewer cases are close; e.g. if X has a bounded density, then $\alpha \geq 1$. Note that the margin condition also leads to faster convergence rates for empirical risk minimization approaches, provided the policy class contains the true classifier. See Audibert and Tsybakov (2007) for further discussion. Under this margin condition, we can further bound the regret of the plug-in nuisance classifier $R_+(\hat{\delta}_+)$, leading to the following corollary to Theorem 4.1.

Corollary 4.3. Under Assumption 2 and the conditions of Theorem 4.1, using the plug-in nuisance classifier $\hat{\delta}_+(x) = \mathbb{1}\{\hat{m}(1,x)+\hat{m}(0,x) \geq 1\}$, the excess worst-case regret of $\hat{\pi}$ relative to

π^* satisfies

$$R_{\text{sup}}(\hat{\pi}, \varpi) - R_{\text{sup}}(\pi^*, \varpi) \leq 2U \times \left(\frac{6+\eta}{\eta} \times \left(2\mathcal{R}_n(\Pi) + \frac{t}{\sqrt{n}} \right) + \|\hat{\gamma} - \gamma\|_2 \sum_{w=0}^{1} \|\hat{m}(w, \cdot) - m(w, \cdot)\|_2 \right)$$
$$+ (u_g - u_l) \times \left(2^{1+\alpha} C \|\hat{m} - m\|_\infty^{1+\alpha} + \frac{t}{2\sqrt{n}} \right),$$

with probability at least $1 - 2\exp\left(-\frac{t^2}{2}\right)$, where $\|\hat{m} - m\|_\infty \equiv \sup_{w,x} |\hat{m}(w, x) - m(w, x)|$, and U is a constant depending on the utility values.

With the plug-in nuisance classifier, $R(\hat{\delta}_+)$ is controlled by the error in the outcome model; however for $\alpha > 0$ this error will be raised to a higher power, leading to a faster rate. In Appendix H, we show an analogous result for the minimax policy relative to the oracle using plug-ins for all nuisance classifiers. See D'Adamo (2023) for an application of these techniques for policy learning in a different partial identification setting, and Kallus (2022) for an application to estimate bounds on $\Pr(Y(1) < Y(0))$.

Finally, although the minimax loss policies we consider are designed to minimize the worst-case expected utility loss, in some cases it may be possible that the true, unidentifiable expected utility loss may also be small. In Appendix E, we conduct a brief simulation study to inspect how the misclassification rates and the true expected utility loss behave in finite samples.

5 Application to Right Heart Catheterization

We now apply the proposed methodology to a particular decision problem: whether or not to use Right Heart Catheterization (RHC) in a clinical setting. RHC is a diagnostic tool where a catheter is inserted into the pulmonary artery. In a controversial observational study, Connors et al. (1996) found that RHC led to an increase in mortality on average. RHC, however, can lead to life-saving treatment for some patients. In this section, we will use the data from Connors et al. (1996) to learn policies for using RHC for certain patients, inspecting how asymmetry in the policy maker's utility function can lead to different data driven decision-making processes.

5.1 Data and setup

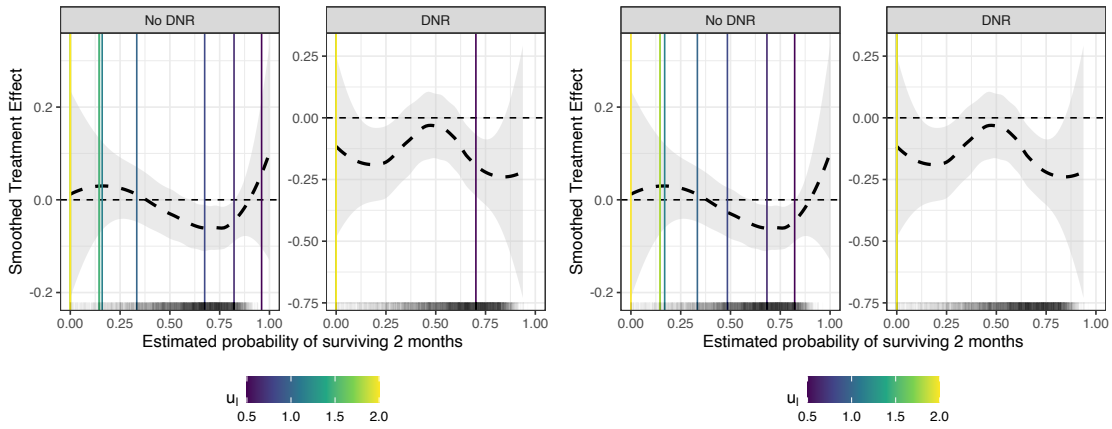
The data from Connors et al. (1996) include $n = 5,735$ ICU patients, 2,184 of whom had RHC applied. We will code the outcome $Y(d) = 1$ as survival by thirty days. In this case, the utility value u_g represents the utility gain in saving a patient's life under RHC who would otherwise die without RHC, and u_l represents the cost of RHC leading to the death of a patient who would otherwise survive. In this study, RHC use was not experimentally randomized and so we will be relying on Assumption 1, using the same set of socioeconomic and health characteristics as those used by Hirano and Imbens (2001) in their propensity score-based analysis.

Throughout, we will use the estimated doubly robust score $\widehat{\Gamma}_w^{\text{dr}}$. To do so, we need estimates of the conditional expectation function $m(w, \cdot)$ and the propensity score $d(\cdot)$. We use a three-fold cross-fitting procedure to estimate these conditional expectations that we detail in Appendix F.1. With the combined DR scores, we estimate that RHC leads to an overall increase in 30-day mortality by 4.5 percentage points, with an estimated standard error of 1.2 percentage points. This result is consistent with the findings of other existing analyses.

To fit each of these models, we use the full set of socioeconomic and health characteristics in the data. We will consider, however, decision rules that only use a subset of the covariates $\mathcal{V} \subset \mathcal{X}$. As we outline in Appendix F.2, Theorem 3.1 implies that the worst-case utility loss will involve nuisance classifiers based on the covariates V, using $m_{\mathcal{V}}(w, v) \equiv \mathbb{E}[Y(w) \mid V = v]$, the expected potential outcome given only the covariates V. To adjust for confounding, however, we still require the DR-scores using the full set of covariates. To construct plug-in estimates of the nuisance classifiers, we use a variant of the DR-learner (Kennedy, 2022), regressing the DR scores $\widehat{\Gamma}_w^{\text{dr}}$ on the covariates V using gradient boosted decision stumps.

5.2 Threshold decision rules with two variables

We begin by considering decision rules that only use two clinical variables: the estimated probability of surviving two months and whether the patient has a Do-Not-Resuscitate (DNR) order.



(a) Relative to always or never using RHC

(b) Relative to oracle policy

Figure 2: Minimax threshold decision rules relative to (a) using RHC for all patients or no patients and (b) the oracle policy as u_l varies and $u_g = 1$ using two variables: the estimated probability of survival and whether a patient has a Do-Not-Resuscitate (DNR) order. The rules assign RHC if the estimated probability of survival is below the threshold. The results are shown separately for patients with (right) and without (left) an DNR order. The dashed line and shaded area are the smoothed estimate of the conditional average treatment effect and 95% confidence interval, respectively.

Throughout we will use threshold decision rules that assign RHC via a cutoff on the estimated probability of surviving two months, using separate thresholds for DNR and non-DNR patients.

First, we consider minimizing the worst-case expected utility loss relative to using RHC for all patients or never using RHC. We estimate the nuisance classifier classifier $\hat{\delta}_+$ via the plug-in approach (Appendix Figure G.1 shows the resulting classifier). We then create estimates of the weighting and cost functions $\hat{c}_0^{\mathbb{1}}(\cdot), \hat{c}_1^{\mathbb{1}}(\cdot), \hat{c}^{\mathbb{1}}(\cdot)$ and solve Eqn (10) to estimate the minimax policy $\hat{\pi}_{\mathbb{1}}$ relative to never using RHC (when $u_g < u_l$) and always using RHC (when $u_g \geq u_l$). We set $u_g = 1$ and vary $u_l \in [0.5, 2]$ so that the utility loss from a harmful treatment moves between half and twice as large as the utility loss from failing to give useful treatment.

Figure 2a shows the resulting decision rules for patients with (right) and without (left) a DNR order. We also estimate the CATE conditioned on the estimated probability of survival and DNR status with the DR-learner using kernel smoothing (Kennedy, 2022). Note that the estimated CATE is positive for non-DNR patients with less than a 50% or greater than 80% probability of survival. Because we restrict to a single threshold, in the symmetric case, this leads to a decision

rule that applies a threshold of 50% for non-DNR patients, while never assigning RHC to DNR patients. As the utility gain in saving a life becomes greater than the cost of causing death, the estimated threshold increases, leading to a decision rule that uses RHC for non-DNR patients with a higher estimated probability of surviving. Eventually the asymmetry is so large in favor of prioritizing useful treatment that almost all non-DNR patients and most DNR patients would be given RHC, even though the CATE is negative. Conversely, as avoiding harm becomes more important, the decision threshold lowers, assigning RHC for fewer and fewer patients until no patients would receive it.

Next, we consider finding the minimax regret policy relative to the oracle $\hat{\pi}_o$, using plug-in estimates of the classifiers $\hat{\delta}_+$, $\hat{\delta}_\tau$, $\hat{\pi}_{\mathbb{O}}$, and $\hat{\pi}_{\mathbb{1}}$.[3] Figure 2b shows the decision functions. Similar to the minimax loss policy relative to always or never using RHC, as u_l decreases the threshold for non-DNR patients increases and as u_l increases the threshold decreases. Even in the extreme case with $u_l = 0.5$, however, DNR patients are not assigned RHC. When $u_l = 2$, DNR patients with a low probability of survival are still assigned RHC. Mirroring our discussion in Section 3.3, measuring regret relative to the best possible policy leads to a less aggressive decision rule than measuring expected utility loss relative to always using RHC.

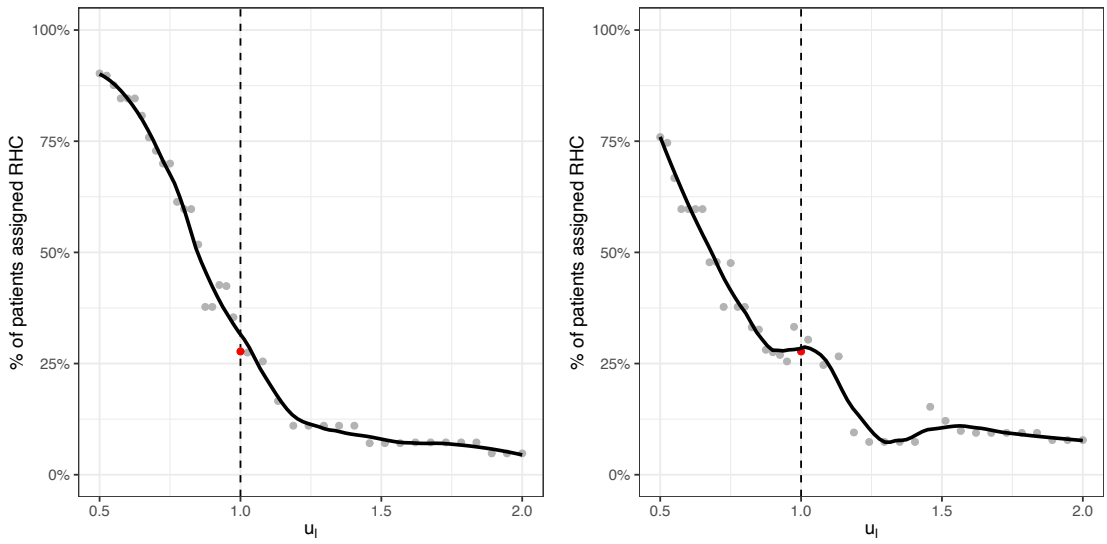
5.3 Decision trees with several clinical variables

Next we move to decision rules with several clinical variables. Recall that Theorem 3.1 shows how to cast the minimax problem as a weighted policy learning problem; so we can find policies from data by solving Eqn (10) using off-the-shelf policy optimization solvers. Here, we focus on learning depth-3 decision trees, using the `policytree` package (Sverdrup et al., 2020).

Because finding the optimal decision tree scales super-linearly with the number of covariates, we first select variables from the set of clinical covariates [4] by fitting a CATE model given the

[3] Note that we use plug-ins for $\hat{\pi}_{\mathbb{O}}$ and $\hat{\pi}_{\mathbb{1}}$ rather than the simple threshold decision rules above, to try to capture the best possible *unconstrained* classifiers rather than the best possible constrained ones.

[4] Of the 66 covariates, 56 are clinical variables while the remaining are socioeconomic variables important to controlling for confounding. See Hirano and Imbens (2001), Table 1, for a full list of covariates.



(a) Relative to always or never using RHC

(b) Relative to oracle policy

Figure 3: Percent of patients assigned RHC under minimax decision rules relative to (a) always or never using RHC and (b) the oracle policy, with $u_l \in [0.5, 2]$ and $u_g = 1$. The line is a smoothed fit. In both panels, the symmetric policy is highlighted in red.

clinical covariates using the DR-learner with random forest regression. We then measure variable importance as the proportion of times a covariate is split on in the forest, weighted by the node depth, using the `grf` package, and select the top 10 most important covariates. See Appendix Figure G.2 for the variable importance measures for all clinical covariates. As before, we consider estimating minimax loss policies relative to always or never using RHC as well as the minimax regret policy relative to the oracle, as the utility asymmetry changes with $u_g = 1$ and $u_l \in [0.5, 2]$. We again use plug-in estimates for the nuisance classifiers with the 10 selected covariates.

Figure 3 shows the percent of patients assigned RHC under the different decision rules. As we move away from the symmetric case towards prioritizing using RHC for patients that will benefit from it, the minimax loss policies relative to always using RHC and to the oracle assign more patients to RHC. In the other direction, as we increase u_l relative to u_g and so seek to prevent harming patients, the minimax loss policies relative to never using RHC and the oracle assign fewer patients RHC. However, the minimax policy relative to the oracle is less extreme, consistent with the two-covariate case in Figure 2 and our discussion in Section 3.3.2.

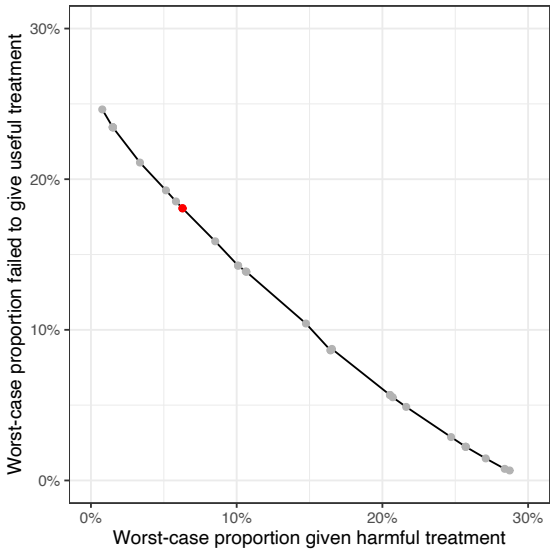


Figure 4: Upper bounds on the proportion of patients that do not receive a useful treatment — the false negative rate — and the proportion of patients given a harmful treatment — the false positive rate — for minimax depth-3 tree policies relative to always or never using RHC, as u_l varies in $[0.5, 2]$ and $u_g = 1$. The red point is the symmetric depth-3 tree policy.

We can measure the impact of these policies in terms of directly interpretable patient outcomes rather than the expected utility loss, by inspecting (a) the worst-case proportion of patients that are given a harmful treatment; (b) the worst-case proportion of patients that are failed to be given a useful treatment; and (c) the overall expected mortality. Following the argument in Section 3, we can find upper bounds on the first two proportions (presented in Appendix H), and use the DR scores $\widehat{\Gamma}_w^{\text{dr}}$ and the plug-in classifier $\hat{\delta}_+$ to get plug-in estimates of them.

Figure 4 shows the worst-case proportion of patients for whom π fails to give a useful treatment (y-axis) versus the worst-case proportion for whom π gives a harmful treatment (x-axis) as we vary $u_l \in [0.5, 2]$. We observe the trade-off between these two types of errors, in the worst-case. If π treats almost no one, in the worst-case almost no patients receive a harmful treatment, but π could be failing to help up to 25% of patients. Conversely, if π treats almost everyone, then π necessarily treats almost everyone for whom it is helpful, but could be providing a harmful treatment for up to 30% of patients. Figure 4 shows the intermediate points between these two extreme scenarios, corresponding to a Pareto frontier between the errors in the worst case.

We can also see the impact on overall mortality along the frontier. Appendix Figure G.3 plots the estimated expected mortality under each policy against the worst-case proportion of patients for whom π fails to give a useful treatment or gives a harmful treatment as we vary $u_l \in [0.5, 2]$. At one extreme, the policy that uses RHC for a large number of patients has the highest expected mortality, because on average RHC is harmful, but the upper bound guarantees that it fails to give a useful treatment in almost no cases. The symmetric policy is at the other end of this tradeoff, with a lower expected mortality but potentially failing to use RHC when it is useful for over 18% of patients. On the other extreme, the policy that rarely uses RHC has a lower expected mortality than always using RHC, but a higher one than the symmetric policy. While the symmetric policy reduces mortality, up to 6% of patients will receive RHC even though it is harmful for them.

6 Discussion

In this paper, we developed a general policy learning framework that allows for asymmetric counterfactual utilities, reflecting common ethical principles including the Hippocratic oath. The asymmetry of utility functions leads to the unidentifiability of expected utilities. We addressed this problem by employing a partial identification approach and then finding a policy that minimizes the maximum expected utility loss relative to a particular policy. We illustrated this framework by reanalyzing the study of Right Heart Catheterization, finding that the minimax optimal policy varies substantially with the asymmetry in the utility and choice of reference policy.

There are several avenues for future research. First, the optimization problem in Eqn (5) is a form of *distributionally robust optimization* (see Bertsimas et al., 2011, for a review). Distributionally robust procedures have been used for risk minimization and policy learning, often by assuming that the true, unknown underlying distribution is close to some known reference distribution. (see e.g., Duchi and Namkoong, 2021; Kallus and Zhou, 2021; Bertsimas et al., 2023). In contrast, Eqn (5) considers all potential joint distributions between the potential outcomes—i.e., all potential principal scores—that agree with the point-identifiable marginal distributions,

without making additional assumptions. A direction for future work is to explicitly encode distributional assumptions. For example, we could treat the case where the potential outcomes are independent as a reference distribution, and assume that the true joint distribution is close to it.

Second, we have restricted our attention to binary outcomes and binary treatments. However, many decision problems involve multiple potential actions and categorical or continuous outcomes. This leads to more principal strata and potential asymmetries in the utility function. We briefly discuss extensions to, and difficulties with, the continuous outcome case in Appendix I, and leave a more thorough analysis to future work. Third, we may consider decision problems with utility functions that also depend on other post-treatment variables or mediators, leading to a different principal stratification structure. Finally, we can consider further constraints that encode notions of fairness, such as the concept of principal fairness that is based on on the principal strata (Imai and Jiang, 2023).

Supplement to "Policy Learning with Asymmetric Counterfactual Utilities"

A Constrained optimization formulation

While we have arrived at the objective defined in Eqn (2) through a utility-based framework, we can also characterize this decision problem in the following constrained form,

$$\begin{aligned} \min_{\pi} \ & \Pr(Y(\pi(X)) < Y(1)) \\ \text{subject to } & \Pr(Y(\pi(X)) < Y(0)) \leq \delta, \\ & \mathbb{E}[\pi(X)] \leq B, \end{aligned} \tag{A.1}$$

where $\Pr(Y(\pi(X)) < Y(0)) = \Pr(Y(1) = 0, Y(0) = 1, \pi(X) = 1)$ and $\Pr(Y(\pi(X)) < Y(1)) = \Pr(Y(1) = 1, Y(0) = 0, \pi(X) = 0)$ represent the probabilities that policy π gives a harmful treatment or fails to give a useful treatment for a randomly selected member of the population, respectively.

In this formulation, the goal is to find a policy π that minimizes the expected proportion of *false negatives* — failing to give a useful treatment — subject to a constraint on the expected proportion of *false positives* — providing a harmful treatment — and the treatment *budget* — the proportion of units treated. Thus, the decision problem given in Eqn (A.1) allows the policy maker to explicitly state their preferences via the constraint on the number of false positives and the budget, rather than implicitly through the utility function in $R_{e_{01}}(\pi, \varpi)$. It is also possible to swap the constraints and the objective to minimize the proportion of false positives subject to a constraint on the proportion of false negatives. We can also interpret Eqn (A.1) through the lens of multiple testing, for each unit i we have a null hypothesis $H_{0i} : Y_i(1) < Y_i(0)$, i.e. that unit i is harmed by treatment. We can view the policy $\pi(X_i)$ as determining whether to reject H_{0i} or not. Then, the constraint on the proportion of false positives in Eqn (A.1) is a scaling of the false detection rate, where the budget constraint limits the number of rejections, and the objective is a measure of the average power under the alternative $H_{1i} : Y_i(1) > Y_i(0)$.

However, note that $\Pr(Y(\pi(X)) < Y(0)) = \mathbb{E}[\pi(X)e_{01}(X)]$ and $\Pr(Y(\pi(X)) < Y(1)) = \mathbb{E}[(1 - \pi(X))(\tau(x) + e_{01}(X))]$. Thus, we can view the expected utility loss $R_{e_{01}}(\pi, \varpi)$ for a constant comparison policy — either always or never providing treatment — as a Lagrangian relaxation of the decision problem defined in Eqn (A.1), where some choice of the false-positive constraint δ and budget B will correspond to a particular value of the utility ratio $\frac{u_g - u_l}{u_g}$ and cost ratio $\frac{c}{u_g}$. This is in contrast to the regret relative to the oracle policy that maximizes the true value, which involves unidentifiable terms in the relative weights on $\tau(x)$ and $e_{01}(x)$, so it cannot be written as a Lagrangian relaxation of Eqn (A.1).

B Connection to maximin policies

Under the maximin approach, we find a policy π that maximizes the worst-case expected utility. In this appendix we connect the minimax loss policies relative to never and always treating to maximin policies under particular choices of the utility. To do so, we need to specify the utilities under no treatment, $u(0; y_1, y_0)$. We consider two cases.

First, say that $u(0; y_1, y_0) = 0$ for all principal strata y_1, y_0. In that case, the expected utility is

$$V(\pi) = \mathbb{E}\left[\pi(X)\left\{u_g\tau(X) + (u_g - u_l)e_{01}(X) - c\right\}\right] = -R_e(\pi, \pi_{\mathbb{O}}).$$

Therefore the maximin policy is equivalent to the minimax loss policy relative to never treating, $\pi^*_{\mathbb{O}}$.

Alternatively, say that the utility function under no treatment mirrors that under treatment, i.e.,

$$u(0; 0, 0) = u(0; 1, 1) = 0, \qquad u(0; 0, 1) = u_l, \qquad u(0; 1, 0) = -u_g.$$

In this case, the expected utility is

$$V(\pi) = \mathbb{E}\left[(\pi(X) - 1)\left\{u_g\tau(X) + (u_g - u_l)e_{01}(X) - c\right\}\right] - c = -R_e(\pi, \pi_{\mathbb{1}}) - c.$$

So, the maximin policy is equivalent to the minimax loss policy relative to always treating, $\pi^*_{\mathbb{1}}$.

C Algorithms for learning minimax loss policies when estimating nuisance functions via empirical risk minimization

Algorithm 1 Estimated minimax policy $\hat{\pi}$ relative to the always-treat policy $\pi^{\mathbb{1}}$ (when $u_g \geq u_l$) and the never-treat policy $\pi^{\mathbb{O}}$ (when $u_g < u_l$)

Input: Policy classes Π and Δ_+

Output: Estimated minimax policy $\hat{\pi}$ relative to $\pi^{\mathbb{1}}$ or $\pi^{\mathbb{O}}$

1: Find $\hat{\delta}_+$ by solving

$$\min_{\delta \in \Delta_+} -\frac{1}{n}\sum_{i=1}^{n} \delta(X_i)\left\{\widehat{\Gamma}_1(X_i, D_i, Y_i) + \widehat{\Gamma}_0(X_i, D_i, Y_i) - 1\right\}.$$

2: Compute weighting and cost functions

$$\hat{c}_1^{\varpi}(x) = u_g + \hat{\delta}_+(x)(u_l - u_g),\ \hat{c}_0^{\varpi}(x) = -u_l - \hat{\delta}_+(x)(u_g - u_l) \text{ and } \hat{c}^{\varpi}(x) = \hat{\delta}_+(x)(u_g - u_l).$$

3: Find a policy $\hat{\pi} \in \arg\min_{\pi \in \Pi} \hat{R}_{\text{sup}}(\pi, \varpi)$.

Algorithm 2 Empirical minimax policy $\hat{\pi}$ relative to the oracle policy π^o

Input: Policy classes Π, Π', Δ_+, and Δ_τ

Output: Empirical minimax policy $\hat{\pi}$ relative to the oracle π^o

1: Find $\hat{\delta}_+$ by solving

$$\min_{\delta \in \Delta_+} -\frac{1}{n} \sum_{i=1}^{n} \delta(X_i) \left\{ \widehat{\Gamma}_1(X_i, D_i, Y_i) + \widehat{\Gamma}_0(X_i, D_i, Y_i) - 1 \right\}.$$

2: **if** $u_g \geq u_l$ **then**

3: Find $\hat{\pi}_{\mathbb{1}}$ via Algorithm 1 with policy class Π'.

4: Find $\hat{\pi}_{\mathbb{O}}$ by solving

$$\min_{\pi \in \Pi'} -\frac{1}{n} \sum_{i=1}^{n} \pi(X_i) \left[u_g \left\{ \widehat{\Gamma}_1(X_i, D_i, Y_i) - \widehat{\Gamma}_0(X_i, D_i, Y_i) \right\} - c \right].$$

5: **else**

6: Find $\hat{\pi}_{\mathbb{O}}$ via Algorithm 1 with policy class Π'.

7: Find $\hat{\pi}_{\mathbb{1}}$ by solving

$$\min_{\pi \in \Pi'} -\frac{1}{n} \sum_{i=1}^{n} \pi(X_i) \left[u_g \left\{ \widehat{\Gamma}_1(X_i, D_i, Y_i) - \widehat{\Gamma}_0(X_i, D_i, Y_i) \right\} - c \right].$$

8: **end if**

9: Find $\hat{\delta}_\tau$ by solving

$$\min_{\delta \in \Delta_\tau} -\frac{1}{n} \sum_{i=1}^{n} \delta(X_i) \left\{ \widehat{\Gamma}_1(X_i, D_i, Y_i) - \widehat{\Gamma}_0(X_i, D_i, Y_i) \right\}.$$

10: Compute weighting and cost functions $\hat{c}_1^{\pi^o}(x), \hat{c}_0^{\pi^o}(x), \hat{c}^{\pi^o}(x)$ via Theorem 3.1.

11: Find the empirical minimax policy $\hat{\pi} \in \arg\min_{\pi \in \Pi} \hat{R}_{\text{sup}}(\pi, \pi^o)$.

D Asymmetric utilities based on observed outcomes

Although it is possible to construct asymmetric utilities without relying on principal strata (Babii et al., 2021), doing so places additional restrictions on the structure of utilities. Consider the following utility function based on observed outcomes alone, $u(d, Y(d)) = u_{11} d Y_i(d) + u_{10} d\{1 - Y_i(d)\} + u_{01}(1-d)Y(d) + u_{00}(1-d)\{1 - Y_i(d)\}$. This utility function includes the interaction between the decision and the observed outcome. Indeed, for a binary decision and outcome, this represents the most general utility that could be specified using the observed outcome.

Table D.1 summarizes the utility gain/loss for treating a unit that belongs to each principal stratum under this setting. With an interaction term, this utility has different utility gains/losses in principal strata $(Y(1) = 1, Y(0) = 0)$ and $(Y(1) = 0, Y(0) = 1)$, allowing for the asymmetry in the utilities as required by the Hippocratic principle. This utility, however, still places restrictions

	$Y_i(0) = 1$	$Y_i(0) = 0$
$Y_i(1) = 1$	Harmless $u_{11} - u_{01}$	Useful $u_{11} - u_{00}$
$Y_i(1) = 0$	Harmful $u_{10} - 2u_{01}$	Useless $u_{10} - u_{01} - u_{00}$

Table D.1: Asymmetric utilities gain/loss for treating a unit, $u(1, Y_i(1)) - u(0, Y_i(0))$ based on the observed outcomes for each of the principal strata. The utility function is given by $u(d, Y_i(d)) = u_{11} d Y_i(d) + u_{10} d\{1 - Y_i(d)\} + u_{01}(1-d) Y_i(d) + u_{00}(1-d)\{1 - Y_i(d)\}$. Each cell corresponds to the principal stratum defined by the values of the two potential outcomes, $Y_i(1)$ and $Y_i(0)$. Each entry represents the utility gain/loss of treatment assignment, relative to no treatment, for a unit that belongs to the corresponding principal stratum.

on its structure. In particular, it requires that the difference between the utility gains in principal strata $(Y(1) = 1, Y(0) = 1)$ and $(Y(1) = 0, Y(0) = 1)$ is the same as that between the utility losses in principal strata $(Y(1) = 1, Y(0) = 0)$ and $(Y(1) = 0, Y(0) = 0)$. Therefore, it might be violated if the difference between harmful and harmless decisions is much greater than that between useful and useless decisions. Thus, a fully general construction of asymmetric utilities requires the use of principal strata, and defining the utility function based on both potential outcomes, $u(d; Y(1), Y(0))$, with utility functions like the one above as a special case.

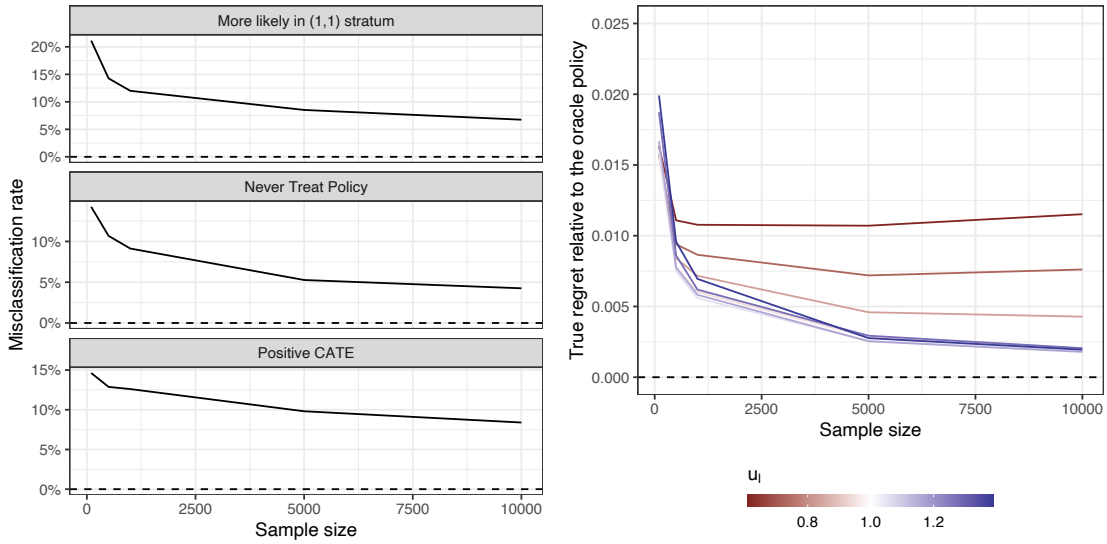
E Simulation study

As the results in Section 4.2 show, the misclassification rates of the underlying nuisance classifiers are important in controlling the excess regret due to estimating the weighting and cost functions that make up the worst-case regret. Additionally, although the minimax policies we consider are designed to minimize the worst-case regret, in some cases it may be possible that the true, unidentifiable regret may also be small. To inspect how the misclassification rates and the true regret behave in finite samples as the sample size increases, we now conduct a brief simulation study, where we can know the true values of the principal scores $e_{y_1 y_0}(x)$.

We first generate n 1-dimensional i.i.d. covariates $X_i \sim N(0, 2)$. We then construct log-linear principal scores as

$$e_{y_1 y_0}(x) = \frac{\exp\left(\alpha_{y_1 y_0} + x\beta_{y_1 y_0}\right)}{\sum_{y_1'=0}^{1} \sum_{y_0'=0}^{1} \exp\left(\alpha_{y_1' y_0'} + x\beta_{y_1' y_0'}\right)},$$

where $(\alpha_{00}, \alpha_{10}, \alpha_{01}, \alpha_{11}) = (.2, .15, 0, 0)$, $\beta_{y_1 y_0} \sim N(0, 40)$ for $(y_1, y_0) \in \{(0,0), (1,0), (0,1)\}$, and $\beta_{11} = 0$. We then jointly sample potential outcomes $\{Y_i(1), Y_i(0)\}$ according to the principal



(a) Misclassification rate for nuisance classifiers (b) Regret of minimax policy relative to oracle

Figure E.1: Performance of nuisance classifiers and the minimax optimal policy relative to the oracle across simulation runs. Panel (a) shows the misclassification rate for the nuisance classifiers $\hat{\delta}_+$ ("More likely in (1,1) stratum") and $\hat{\delta}_\tau$ ("Positive CATE"), as well as the minimax policy relative to the never-treat policy for $u_l = 0.833$. Panel (b) shows the true regret of the minimax optimal policy relative to the oracle, in the sample, for $u_g = 1$ and as u_l varies between 0.5 and 1.5.

scores at covariate value X_i. In this simulation study, we consider a randomized control trial with binary treatment D_i sampled independently as Bernoulli random variables with probability one half.

For each value of sample size $n \in \{100, 500, 1000, 5000, 10000\}$, we draw 1,000 samples according to the above data generating process. In each simulation run, we find the minimax optimal policy with respect to the oracle following Algorithm 2 with zero cost $c = 0$, $u_g = 1$, and u_l varying between 0.6 and 1.4, where the value of u_l changes within each simulation run.

We use the IPW scoring function and restrict all policy classes to be the set of linear thresholds, solving the optimization problem exactly by direct search. Figure E.1a shows the average misclassification rate for the nuisance classifiers $\hat{\delta}_+$ and $\hat{\delta}_\tau$, as well as the misclassification rate for the the minimax policy relative to always treating for $u_l = 0.833$. As we expect, we see that these misclassification rates decrease as the sample size increases.

Figure E.1b shows the true regret of the minimax policy relative to the oracle as u_l varies. Since the oracle is the best possible policy, this regret is always positive. The regret does decrease along with the sample size, reflecting both the decrease in the nuisance misclassification rate and the decrease in the worst-case excess regret when the nuisance classifiers are known. Notice, however, that the regret does stop decreasing after a certain point, flattening out at a different level depending on the asymmetry in the utility function. In highly asymmetric settings where u_ℓ

is small the regret is essentially flat. This is due to the fundamental identifiability problem, and even with infinite data we cannot guarantee that the true regret will be zero. In contrast, in the symmetric setting the regret continues to decrease as the sample size increases.

F Implementation details for application to RHC

F.1 Details on cross-fitting procedure

In the empirical application to Right Heart Catheterization in Section 5, we use a three-fold cross fitting procedure to estimate the nuisance functions. We then use the plug-in method to estimate the nuisance classifiers. Below we present this procedure step-by-step

1. Randomly split the data into three folds.

2. For each fold $k = 1, 2, 3$, estimate the outcome model $\hat{m}^{-k}(\cdot, \cdot)$ and $\hat{d}^{-k}(\cdot)$ on the two other folds via gradient boosted decision stumps.

3. For each unit i, denote $k[i]$ as the fold that it belongs to, then obtain estimates of the outcome model $\hat{m}^{-k[i]}(w, X_i)$, the propensity score $\hat{d}_w^{-k[i]}(X_i)$, and the IP weight $\hat{\gamma}_w^{-k[i]}(D_i, X_i)$.

4. Use these to construct cross-fit estimates of the DR scoring rule:

$$\widehat{\Gamma}_w^{-k[i]}(X_i, D_i, Y_i) = \hat{m}^{-k[i]}(w, X_i) + \{Y_i - \hat{m}^{-k[i]}(w, X_i)\}\hat{\gamma}_w^{-k[i]}(X_i, D_i),$$

and cross-fit plug-in estimates of the classifiers

$$\hat{\delta}_+^{-k[i]}(X_i) = \mathbb{1}\{\hat{m}^{-k[i]}(1, X_i) + \hat{m}^{-k[i]}(0, X_i) \geq 1\},$$
$$\hat{\delta}_\tau^{-k[i]}(X_i) = \mathbb{1}\{\hat{m}^{-k[i]}(1, X_i) - \hat{m}^{-k[i]}(0, X_i) \geq 0\},$$
$$\hat{\pi}_{\mathbb{O}}^{-k[i]}(X_i) = \begin{cases} \mathbb{1}\left\{\hat{m}^{-k[i]}(1, X_i) - \hat{m}^{-k[i]}(0, X_i) \geq \frac{c}{u_g}\right\}, & u_g \geq u_l, \\ \mathbb{1}\left\{\hat{m}^{-k[i]}(1, X_i) \geq \frac{u_l}{u_g}\hat{m}^{-k[i]}(0, X_i) + \frac{c}{u_g}\right\}, & u_g < u_l \text{ and } \hat{\delta}_+^{-k[i]}(X_i) = 0, \\ \mathbb{1}\left\{\hat{m}^{-k[i]}(1, X_i) \geq \frac{u_g}{u_l}\hat{m}^{-k[i]}(0, X_i) + \frac{u_l - u_g + c}{u_l}\right\}, & u_g < u_l \text{ and } \hat{\delta}_+^{-k[i]}(X_i) = 1, \end{cases}$$
$$\hat{\pi}_{\mathbb{1}}^{-k[i]}(X_i) = \begin{cases} \mathbb{1}\left\{\hat{m}^{-k[i]}(1, X_i) - \hat{m}^{-k[i]}(0, X_i) \geq \frac{c}{u_g}\right\}, & u_g < u_l, \\ \mathbb{1}\left\{\hat{m}^{-k[i]}(1, X_i) \geq \frac{u_l}{u_g}\hat{m}^{-k[i]}(0, X_i) + \frac{c}{u_g}\right\}, & u_g \geq u_l \text{ and } \hat{\delta}_+^{-k[i]}(X_i) = 0, \\ \mathbb{1}\left\{\hat{m}^{-k[i]}(1, X_i) \geq \frac{u_g}{u_l}\hat{m}^{-k[i]}(0, X_i) + \frac{u_l - u_g + c}{u_l}\right\}, & u_g \geq u_l \text{ and } \hat{\delta}_+^{-k[i]}(X_i) = 1. \end{cases}$$

Then plug in the cross-fit classifiers into the formulas in Appendix H to create cross-fit estimates of $\hat{c}^{-k[i]\varpi}(X_i)$.

5. Solve Eqn (10) with the cross-fit estimates:

$$\hat{\pi} \in \arg\min_{\pi \in \Pi} -\frac{1}{n}\sum_{i=1}^{n} \pi(X_i)\left\{\hat{c}_1^{-k[i]\varpi}(X_i)\widehat{\Gamma}_1^{-k[i]}(X_i, D_i, Y_i) + \hat{c}_0^{-k[i]\varpi}(X_i)\widehat{\Gamma}_0^{-k[i]}(X_i, D_i, Y_i) + \hat{c}^{-k[i]\varpi}(X_i)\right\}$$

F.2 Minimax loss policies using a subset of covariates

It is often that case that we wish to construct minimax loss decision rules that only use a subset of the covariates $\mathcal{V} \subset \mathcal{X}$. To consider this case, define $m_{\mathcal{V}}(w, v) \equiv \mathbb{E}[Y(w) \mid V = v]$ to be the expected potential outcome conditioned on the subset of covariates v. Applying Theorem 3.1 to this setting, we get that we can write the worst-case expected utility loss of π relative to ϖ as

$$R_{\text{sup}}(\pi, \varpi) = C - \mathbb{E}\left[\pi(X)\left\{c^{\varpi}_{1\mathcal{V}}(V) m_{\mathcal{V}}(1, V) + c^{\varpi}_{0\mathcal{V}}(V) m_{\mathcal{V}}(0, V) + c^{\varpi}_{\mathcal{V}}(V)\right\}\right],$$

where the weighting and cost functions $c^{\varpi}_{1\mathcal{V}}(\cdot), c^{\varpi}_{0\mathcal{V}}(\cdot), c^{\varpi}_{\mathcal{V}}(\cdot)$ depend on the nuisance classifiers given only the subset of the covariates V, i.e.

$$\begin{aligned}
\delta_{+\mathcal{V}}(v) &= \mathbb{1}\{m_{\mathcal{V}}(1, v) + m_{\mathcal{V}}(0, v) \geq 1\}, \\
\delta_{\tau\mathcal{V}}(v) &= \mathbb{1}\{m_{\mathcal{V}}(1, v) - m_{\mathcal{V}}(0, v) \geq 0\}, \\
\pi^*_{\mathbb{O}\mathcal{V}}(v) &= \begin{cases}
\mathbb{1}\left\{m_{\mathcal{V}}(1, v) - m_{\mathcal{V}}(0, v) \geq \frac{c}{u_g}\right\}, & u_g \geq u_l, \\
\mathbb{1}\left\{m_{\mathcal{V}}(1, v) \geq \frac{u_l}{u_g} m_{\mathcal{V}}(0, v) + \frac{c}{u_g}\right\}, & u_g < u_l \text{ and } \delta_{+\mathcal{V}}(v) = 0, \\
\mathbb{1}\left\{m_{\mathcal{V}}(1, v) \geq \frac{u_g}{u_l} m_{\mathcal{V}}(0, v) + \frac{u_l - u_g + c}{u_l}\right\}, & u_g < u_l \text{ and } \delta_{+\mathcal{V}}(v) = 1,
\end{cases} \\
\pi^*_{\mathbb{1}\mathcal{V}}(v) &= \begin{cases}
\mathbb{1}\left\{m_{\mathcal{V}}(1, v) - m_{\mathcal{V}}(0, v) \geq \frac{c}{u_g}\right\}, & u_g < u_l, \\
\mathbb{1}\left\{m_{\mathcal{V}}(1, v) \geq \frac{u_l}{u_g} m_{\mathcal{V}}(0, v) + \frac{c}{u_g}\right\}, & u_g \geq u_l \text{ and } \delta_{+\mathcal{V}}(v) = 0, \\
\mathbb{1}\left\{m_{\mathcal{V}}(1, v) \geq \frac{u_g}{u_l} m_{\mathcal{V}}(0, v) + \frac{u_l - u_g + c}{u_l}\right\}, & u_g \geq u_l \text{ and } \delta_{+\mathcal{V}}(v) = 1.
\end{cases}
\end{aligned}$$

However, note that in order to use observable data, we must account for confounding, since in general $m_{\mathcal{V}}(w, v) \neq \mathbb{E}(Y \mid V = v, W = w)$ when $\mathcal{V}$ is a subset of $\mathcal{X}$. We can however, still use the IPW or DR scoring functions since $m_{\mathcal{V}}(w, v) = \mathbb{E}[\Gamma_w(X, D, Y) \mid V = v]$. So we can write the worst-case expected utility loss in terms of the scoring functions—where we condition on X—and the nuisance classifiers only conditioning on the subset of covariates V:

$$R_{\text{sup}}(\pi, \varpi) = C - \mathbb{E}\left[\pi(V)\left\{c^{\varpi}_{1\mathcal{V}}(V)\Gamma_1(X, D, Y) + c^{\varpi}_{0\mathcal{V}}(V)\Gamma_0(X, D, Y) + c\mathcal{V}^{\varpi}(V)\right\}\right],$$

Constructing plug-in estimates of the nuisance classifiers involves estimating $m_{\mathcal{V}}(w, v) = \mathbb{E}[\Gamma_w(X, D, Y) \mid V = v]$, which we can do by regressing the estimated DR scores on the subset of the covariates V, a variant of the DR-learner (Kennedy, 2022).

Overall, this leads to the following steps:

1. Estimate the DR score $\widehat{\Gamma}_w(x, d, y)$ using *all covariates* X to account for confounding.

2. Estimate the expected potential outcomes given the subset of covariates V, $\hat{m}_{\mathcal{V}}(w, v)$ using the DR-learner and regressing the estimates $\widehat{\Gamma}_w(X_i, D_i, Y_i)$ on V.

3. Form plug in estimates of the nuisance classifiers, e.g. $\hat{\delta}_\tau(v) = \mathbb{1}\{\hat{m}_{\mathcal{V}}(1, v) - \hat{m}_{\mathcal{V}}(0, v)\}$

and $\hat{\delta}_+(v) = \mathbb{1}\{\hat{m}_{\mathcal{V}}(1,v) + \hat{m}_{\mathcal{V}}(0,v) - 1 \geq 0\}$.

4. Get plug-in estimates of the weighting and cost functions $\hat{c}^{\varpi}_{1\mathcal{V}}(V_i), \hat{c}^{\varpi}_{0\mathcal{V}}(V_i), \hat{c}^{\varpi}_{\mathcal{V}}(V_i)$, using the estimates of the nuisance classifiers.

5. Find the policy $\hat{\pi} : \mathcal{V} \to \{0,1\}$ by solving

$$\min_{\pi \in \Pi} -\frac{1}{n} \sum_{i=1}^{n} \pi(V) \left\{ \hat{c}^{\varpi}_{1\mathcal{V}}(V) \widehat{\Gamma}_1(X,D,Y) + \hat{c}^{\varpi}_{0\mathcal{V}}(V) \widehat{\Gamma}_0(X,D,Y) + \hat{c}\mathcal{V}^{\varpi}(V) \right\}.$$

Finally, note that as in Section F.1 above, we can use cross-fit estimates here, where for each fold k, both $\widehat{\Gamma}_w^{-k}$ and $\hat{m}_{\mathcal{V}}^{-k}$ are fit on data not in fold k. In principle we could do a multi-stage cross-fitting procedure, where for each fold k, we further break up the fold into sub-folds and cross-fit $\hat{m}_{\mathcal{V}}^{-k}$ within the fold k. We opt to use a simpler cross-fitting procedure here, noting that it may impact the quality of the DR-learner estimate $\hat{m}_{\mathcal{V}}^{-k}$.

G Additional figures

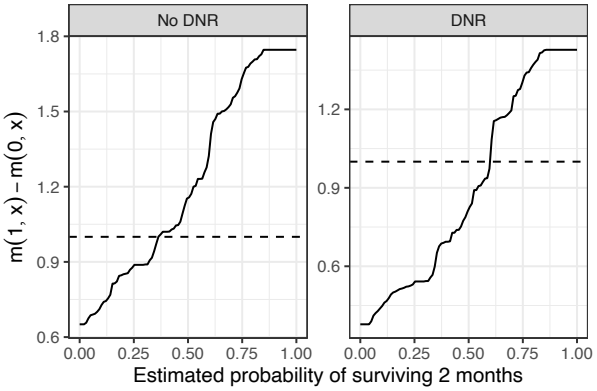


Figure G.1: Plug-in estimate of the decision rule $\hat{\delta}_+(v)$ to classify whether $m_{\mathcal{V}}(1,v) + m_{\mathcal{V}}(0,x) \geq 1$ using the estimated probability of survival and DNR status.

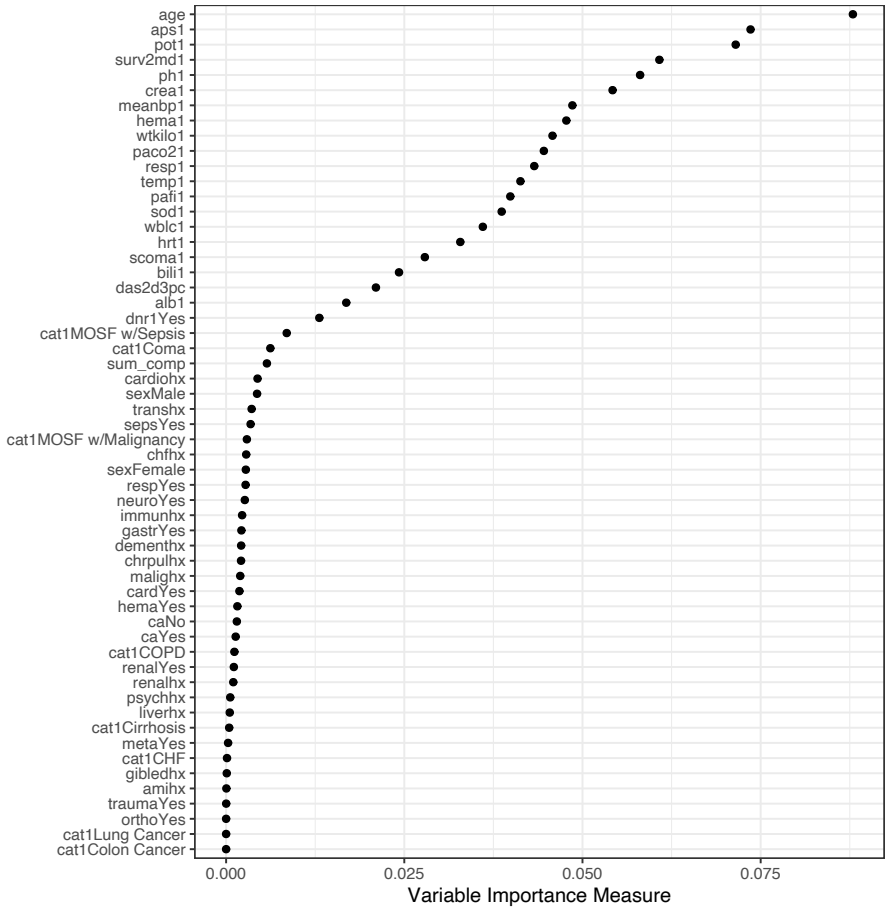


Figure G.2: Variable importance for estimated CATE.

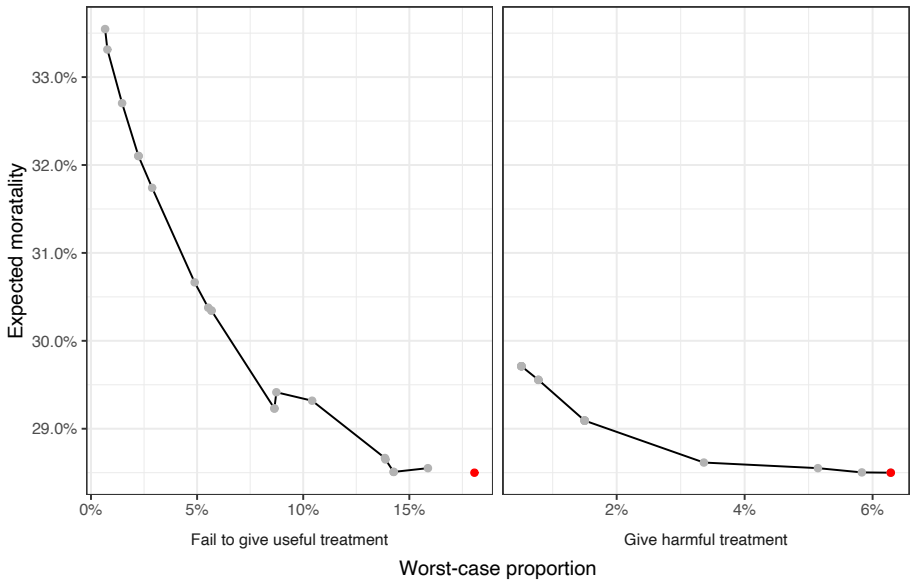


Figure G.3: Estimated expected mortality versus the worst case proportion of cases where the policy fails to give a useful treatment (left panel) and where the policy gives a harmful treatment (right panel), for linear minimax policies relative to never using RHC, as u_l varies in $[0.65, 1.2]$ and $u_g = 1$. The red point is the symmetric linear policy.

H Additional results

Full statement of Theorem 3.1 Let $\pi : \mathcal{X} \to \{0, 1\}$ be a deterministic policy. For comparison policy $\varpi \in \{\pi^{\mathbb{O}}, \pi^{\mathbb{1}}\}$, the worst-case expected utility loss of π relative to ϖ is

$$\begin{aligned} R_{\sup}(\pi, \varpi) &= C - \mathbb{E}\left[\pi(X)\left\{c_1^{\varpi}(X)m(1, X) + c_0^{\varpi}(X)m(0, X) + c^{\varpi}(X)\right\}\right] \\ &= C - \mathbb{E}\left[\pi(X)\left\{c_1^{\varpi}(X)\Gamma_1(X, D, Y) + c_0^{\varpi}(X)\Gamma_0(X, D, Y) + c^{\varpi}(X)\right\}\right], \end{aligned} \tag{H.1}$$

where C is a constant that does not depend on π. For $u_g \geq u_l$,

$$\begin{aligned} c_1^{\pi^{\mathbb{O}}}(x) &= u_l + (u_g - u_l)\delta_\tau(x) & c_0^{\pi^{\mathbb{O}}}(x) &= -u_l - (u_g - u_l)\delta_\tau(x) & c^{\pi^{\mathbb{O}}}(x) &= -c, \\ c_1^{\pi^{\mathbb{1}}}(x) &= u_g + \delta_+(x)(u_l - u_g) & c_0^{\pi^{\mathbb{1}}}(x) &= -u_l - \delta_+(x)(u_g - u_l) & c^{\pi^{\mathbb{1}}}(x) &= \delta_+(x)(u_g - u_l) - c, \end{aligned}$$

and for $u_g < u_l$,

$$\begin{aligned} c_1^{\pi^{\mathbb{O}}}(x) &= u_g + \delta_+(x)(u_l - u_g) & c_0^{\pi^{\mathbb{O}}}(x) &= -u_l - \delta_+(x)(u_g - u_l) & c^{\pi^{\mathbb{O}}}(x) &= \delta_+(x)(u_g - u_l) - c, \\ c_1^{\pi^{\mathbb{1}}}(x) &= u_l + (u_g - u_l)\delta_\tau(x) & c_0^{\pi^{\mathbb{1}}}(x) &= -u_l - (u_g - u_l)\delta_\tau(x) & c^{\pi^{\mathbb{1}}}(x) &= -c. \end{aligned}$$

Define $\pi^*_{\mathbb{O}} \equiv \arg\min_\pi R_{\sup}(\pi, \pi^{\mathbb{O}})$ and $\pi^*_{\mathbb{1}} \equiv \arg\min_\pi R_{\sup}(\pi, \pi^{\mathbb{1}})$ as the minimax expected utility loss solutions relative to the never-treat policy and always-treat policy, respectively. The worst-case regret relative to the oracle policy π^o is of the form in Eqn (H.1) where for $u_g \geq u_l$,

$$\begin{aligned} \begin{pmatrix} c_1^{\pi^o}(x) \\ c_0^{\pi^o}(x) \\ c^{\pi^o}(x) \end{pmatrix} =& (1 - \pi^*_{\mathbb{1}}(x)) \begin{pmatrix} u_l + (u_g - u_l)\delta_\tau(x) \\ -u_l - (u_g - u_l)\delta_\tau(x) \\ -c \end{pmatrix} + \pi^*_{\mathbb{O}}(x) \begin{pmatrix} u_g - (u_g - u_l)\delta_+(x) \\ -u_l - (u_g - u_l)\delta_+(x) \\ (u_g - u_l)\delta_+(x) - c \end{pmatrix} \\ &+ (1 - \pi^*_{\mathbb{O}}(x))\pi^*_{\mathbb{1}}(x) \begin{pmatrix} u_l + u_g + (u_g - u_l)(\delta_\tau(x) - \delta_+(x)) \\ -2u_l - (u_g - u_l)(\delta_\tau(x) + \delta_+(x)) \\ (u_g - u_l)\delta_+(x) - 2c \end{pmatrix}, \end{aligned}$$

and for $u_g < u_l$,

$$\begin{aligned} \begin{pmatrix} c_1^{\pi^o}(x) \\ c_0^{\pi^o}(x) \\ c^{\pi^o}(x) \end{pmatrix} =& (1 - \pi^*_{\mathbb{1}}(x)) \begin{pmatrix} u_g - (u_g - u_l)\delta_+(x) \\ -u_l - (u_g - u_l)\delta_+(x) \\ (u_g - u_l)\delta_+(x) - c \end{pmatrix} + \pi^*_{\mathbb{O}}(x) \begin{pmatrix} u_l + (u_g - u_l)\delta_\tau(x) \\ -u_l - (u_g - u_l)\delta_\tau(x) \\ -c \end{pmatrix} \\ &+ (1 - \pi^*_{\mathbb{O}}(x))\pi^*_{\mathbb{1}}(x) \begin{pmatrix} u_l + u_g + (u_g - u_l)(\delta_\tau(x) - \delta_+(x)) \\ -2u_l - (u_g - u_l)(\delta_\tau(x) + \delta_+(x)) \\ (u_g - u_l)\delta_+(x) - 2c \end{pmatrix}. \end{aligned}$$

Corollary H.1 (Minimax regret relative to the always-treat policy)**.** If $u_g \geq u_l$, the minimax regret solution to Equation (5), $\pi^*_{\mathbb{1}} \equiv \arg\min_\pi R_{\sup}(\pi, \pi^{\mathbb{1}})$, is

$$\pi^*_{\mathbb{1}}(x) = \begin{cases} \mathbb{1}\left\{m(1, x) \geq \frac{u_l}{u_g} m(0, x) + \frac{c}{u_g}\right\}, & \delta_+(x) = 0, \\ \mathbb{1}\left\{m(1, x) \geq \frac{u_g}{u_l} m(0, x) + \frac{u_l - u_g + c}{u_l}\right\}, & \delta_+(x) = 1. \end{cases}$$

Otherwise, if $u_g < u_l$, it is given by the symmetric policy,

$$\pi^*_{\mathbb{1}}(x) = \mathbb{1}\left\{\tau(x) \geq \frac{c}{u_g}\right\} = \pi^{\text{symm}}(x).$$

Assumption H.1. There exists an $\alpha > 0$ and a constant C such that for any $t \geq 0$,

(a) $\Pr(|m(1,X) + m(0,X) - 1| \leq t) \leq Ct^\alpha$.

(b) $\Pr(|m(1,X) - m(0,X)| \leq t) \leq Ct^\alpha$.

(c) For $u_g > u_l$ and c,

$$\Pr\left(|\{u_g - (u_g - u_l)\delta_+(X)\}m(1,X) - \{u_l + (u_g - u_l)\delta_+(X)\}m(0,X) + (u_g - u_l)\delta_+(X) - c| \leq t\right) \leq C$$

(d) For $u_g > u_l$ and c,

$$\Pr(|\{u_l + (u_g - u_l)\delta_\tau(X)\}\tau(X) - c| \leq t) \leq Ct^\alpha.$$

Theorem H.2. Let $u_g \geq u_l$, and define

$$\widehat{L}_b(x) = \{u_l + \hat{\delta}_\tau(x)(u_g - u_l)\}(\hat{m}(1,x) - \hat{m}(0,x)) - c,$$
$$\widehat{U}_b(x) = \{u_g - (u_g - u_l)\hat{\delta}_+(x)\}\hat{m}(1,x) - \{u_l + (u_g - u_l)\hat{\delta}_+(x)\}\hat{m}(0,x) + (u_g - u_l)\hat{\delta}_+(x) - c,$$

and let $\hat{\pi}^{\text{plug}}_{\mathbb{0}}(x) = \mathbb{1}\{\widehat{L}_b(x) \geq 0\}$ and $\hat{\pi}^{\text{plug}}_{\mathbb{1}}(x) = \mathbb{1}\{\widehat{U}_b(x) \geq 0\}$ be the plug-in estimates of the minimax optimal policies relative to never or always treating. Under Assumptions H.1(b) and H.1(d), the excess worst case regret for $\hat{\pi}^{\text{plug}}_{\mathbb{0}}$ relative to $\pi^*_{\mathbb{0}}$ is

$$R_{\text{sup}}(\hat{\pi}^{\text{plug}}_{\mathbb{0}}, \pi^{\mathbb{0}}) - R_{\text{sup}}(\pi^*_{\mathbb{0}}, \pi^{\mathbb{0}}) \leq 2^{1+\alpha}CU\|m - \hat{m}\|^{1+\alpha}_\infty,$$

where U is a constant depending on the utility values, α, and C. Under Assumptions 2 and H.1(c), the excess worst case regret for $\hat{\pi}^{\text{plug}}_{\mathbb{1}}$ relative to $\pi^*_{\mathbb{1}}$ is

$$R_{\text{sup}}(\hat{\pi}^{\text{plug}}_{\mathbb{1}}, \pi^{\mathbb{1}}) - R_{\text{sup}}(\pi^*_{\mathbb{0}}, \pi^{\mathbb{0}}) \leq 2^{1+\alpha}CU\|m - \hat{m}\|^{1+\alpha}_\infty,$$

where U is a constant depending on the utility values, α, and C.

Corollary H.3. Let $u_g \geq u_l$, $\hat{\pi}_o$ be a solution to Equation (10) with alternative policy $\varpi = \pi^o$ and with nuisance functions $\hat{m}$ and $\hat{d}$ fit on a separate sample and nuisance classifiers $\hat{\delta}_+(x) = \mathbb{1}\{m(1,x) + m(0,x) - 1 \geq 0\}, \hat{\delta}_\tau(m(1,x) - m(0,x) \geq 0), \hat{\pi}^{\text{plug}}_{\mathbb{0}}$, and $\hat{\pi}^{\text{plug}}_{\mathbb{1}}$, and let π^*_o be a solution to Equation (5), with alternative policy $\varpi = \pi^o$. Under the strict overlap condition in

Assumption 1, the excess worst-case regret of $\hat{\pi}_o$ relative to π_o^* satisfies

$$R_{\text{sup}}(\hat{\pi}_0, \pi^o) - R_{\text{sup}}(\pi_o^*, \pi^o) \leq 2U_1 \times \left(\frac{6+\eta}{\eta} \times \left(2\mathcal{R}_n(\Pi) + \frac{t}{\sqrt{n}} \right) + \|\hat{m} - m\|_2 \|\hat{\gamma} - \gamma\|_2 \right) + 2^{2+\alpha} C U_2 \|\hat{m} - m\|_\infty^{1+\alpha} + (u_g - u_l) \frac{t}{2\sqrt{n}},$$

with probability at least $1 - 2\exp\left(-\frac{t^2}{2}\right)$, where U_1 is a constant depending on the utility values, and U_2 is a constant depending on the utility values, α, and C.

Upper bounds on worst-case proportion of units given a harmful treatment or are failed to be given a useful treatment.
First, note that

$$\Pr(Y(\pi(X)) < Y(0)) = \Pr(\pi(X) = 1, Y(0) = 1, Y(1) = 0) = \mathbb{E}\left[\pi(X) e_{01}(X)\right],$$
$$\Pr(Y(\pi(X)) < Y(1)) = \Pr(\pi(X) = 0, Y(0) = 0, Y(1) = 1) = \mathbb{E}\left[(1 - \pi(X))(\tau(X) + e_{01}(X))\right].$$

Plugging in the upper and lower bounds on $e_{01}(X)$ in Section 3, we get the following upper bounds:

$$\sup_{e_{01}(x) \in [L(x), U(x)]} \Pr(Y(\pi(X)) < Y(0)) = \mathbb{E}\left(\pi(X)\left[m(0,X) + \delta_+(X)\left\{1 - m(1,X) - m(0,X)\right\}\right]\right),$$
$$\sup_{e_{01}(x) \in [L(x), U(x)]} \Pr(Y(\pi(X)) < Y(1)) = \mathbb{E}\left(\left\{1 - \pi(X)\right\}\left[m(1,X) + \delta_+(X)\left\{1 - m(1,X) - m(0,X)\right\}\right]\right).$$

I Continuous outcomes

Here we briefly consider extending our framework to the case with a binary decision $D \in \{0, 1\}$ but continuous potential outcomes $(Y(0), Y(1)) \in \mathbb{R}^2$. We define the utility function $u(d; y_1, y_0)$ as before and write the value function as

$$V(\pi) = \mathbb{E}\left[u(0; Y(1), Y(0)) + \pi(X) \times (u(1; Y(1), Y(0)) - u(0; Y(1), Y(0)))\right].$$

Defining $e_{y_1 y_0}(x)$ as the conditional joint density of the potential outcomes given $X = x$, the expected utility loss relative to ϖ is

$$V(\varpi) - V(\pi) = \mathbb{E}\left[\pi(X) \int_{y_1} \int_{y_0} (u(1; y_1, y_0) - u(0; y_1, y_0)) e_{y_1 y_0}(x) dy_0 dy_1\right].$$

With continuous outcomes, there are many potential ways to choose the utility function. One option is a utility function such that $u(1; y_1, y_0) - u(0; y_1, y_0) = y_1 - y_0 - u_\ell \mathbb{1}\{y_1 < y_0\}$. This is analogous to the utility function with binary outcomes, with an explicit utility gain/loss associated with a harmful ($Y(1) < Y(0)$) or useful ($Y(1) > Y(0)$) treatment. Defining the conditional probability of harm as $h(x) = \Pr(Y(1) < Y(0) \mid X = x)$, we can write the expected

utility loss as

$$V(\varpi) - V(\pi) = \mathbb{E}\left[\pi(X)\{\tau(x) - u_\ell h(x)\}\right].$$

As in the binary case, we can use sharp bounds on the distribution of individual treatment effects (Fan and Park, 2010), $h(x) \in [L_h(x), U_h(x)]$, where

$$L_h(x) = \max\{\sup_y\{F_1(y \mid x) - F_0(y \mid x)\}, 0\},$$
$$U_h(x) = 1 + \min\{\inf_y\{F_1(y \mid x) - F_0(y \mid x)\}, 0\},$$

where $F_1(\cdot \mid x), F_0(\cdot \mid x)$ are the marginal CDFs conditional on $X = x$ for the potential outcomes under treatment and control, respectively. Now we can again define the minimax expected utility loss policy as the policy that solves

$$\min_\pi \max_{h(x)\in[L(x),U(x)]} \mathbb{E}\left[\pi(X)\{\tau(x) - u_l h(x)\}\right].$$

While this again leads to a point-identifiable objective, we note two ways in which this problem is more difficult than with binary outcomes. First, the upper and lower bounds on the probability of harm involves the conditional CDFs of $Y(1)$ and $Y(0)$. These can be more difficult to estimate than the conditional expected outcomes. Second, the bounds involve supremums and infimums over all $y \in \mathbb{R}$. This may require a more careful analysis and stronger assumptions in order to ensure that the default plug-in approach that we suggest for the binary outcome case will lead to reasonable guarantees on the excess expected utility loss.

J Proofs and derivations

J.1 Main results

Derivation of the expected utility loss First, notice that the expected utility of policy π is

$$V(\pi) = \mathbb{E}\left[\sum_{y_1=0}^{1}\sum_{y_0=0}^{1} e_{y_1y_0}(X)u(0; y_1, y_0)\right] + \underbrace{\mathbb{E}\left[\sum_{y_1=0}^{1}\sum_{y_0=0}^{1} \pi(X)e_{y_1y_0}(X)\left\{u(1; y_1, y_0) - u(0; y_1, y_0)\right\}\right]}_{(*)}.$$

The second term can be written as

$$\begin{aligned}
(*) &= \mathbb{E}\left[\pi(X)\left\{e_{10}(X)(u_g - c) - e_{01}(X)(u_l + c) - e_{00}(X)c - e_{11}(X)c\right\}\right] \\
&= \mathbb{E}\left[\pi(X))\left\{e_{10}(X)u_g - e_{01}(X)u_l - c(e_{10}(X) + e_{01}(X) + e_{00}(X) + e_{11}(X))\right\}\right] \\
&= \mathbb{E}\left[\pi(X)\left\{(\tau(X) + e_{01}(X))u_g - e_{01}(X)u_l - c\right\}\right] \\
&= \mathbb{E}\left[\pi(X)\left\{u_g\tau(X) + (u_g - u_l)e_{01}(X) - c\right\}\right],
\end{aligned}$$

where we have used the fact that $\tau(x) = e_{10}(x) - e_{01}(x)$. So the expected utility loss of policy π relative to policy ϖ is

$$V(\varpi) - V(\pi) = \mathbb{E}\left[(\varpi(X) - \pi(X))\left\{u_g\tau(X) + (u_g - u_l)e_{01}(X) - c\right\}\right].$$

Proof of Theorem 3.1. Define $b(x) = u_g\tau(x) + (u_g - u_l)e_{01}(X) - c$, and

$$\begin{aligned} L_b(x) &= \min_{e(x)\in[L(x),U(x)]} \left\{u_g\tau(x) + (u_g - u_l)e_{01}(X) - c\right\}, \\ U_b(x) &= \max_{e(x)\in[L(x),U(x)]} \left\{u_g\tau(x) + (u_g - u_l)e_{01}(X) - c\right\}. \end{aligned}$$

Note that the worst-case regret relative to the always and never treat policies are

$$\begin{aligned} R_{\text{sup}}(\pi, \pi^{\mathbb{O}}) &= -\mathbb{E}[\pi(X)L_b(X)], \\ R_{\text{sup}}(\pi, \pi^{\mathbb{1}}) &= \mathbb{E}[\{1 - \pi(X)\}U_b(X)] = \mathbb{E}[U_b(X)] - \mathbb{E}[\pi(X)U_b(X)]. \end{aligned}$$

From this, we can find the unconstrained minimax regret policies

$$\begin{aligned} \pi^*_{\mathbb{O}} &= \arg\min_{\pi} -\mathbb{E}[\pi(X)L_b(X)] = \quad \mathbb{1}\{L_b(x) \geq 0\}, \\ \pi^*_{\mathbb{1}} &= \arg\min_{\pi} -\mathbb{E}[\pi(X)U_b(X)] = \quad \mathbb{1}\{U_b(x) \geq 0\}. \end{aligned}$$

Now, the oracle policy is $\pi^o(x) = \mathbb{1}\{b(x) \geq 0\}$. So if $L_b(x) \geq 0 \Leftrightarrow \pi^*_{\mathbb{O}}(x) = 1$ then $\pi^o(x) = 1$ for all possible values of the principal score $e_{01}(x)$. In this case,

$$\max_{e(x)\in[L(x),U(x)]} \{\pi^o(x) - \pi(x)\}b(x) = \{1 - \pi(x)\}U_b(x).$$

Similarly, if $U_b(x) < 0 \Leftrightarrow \pi^*_{\mathbb{1}}(x) = 0$ then $\pi^o(x) = 0$, and

$$\max_{e(x)\in[L(x),U(x)]} \{\pi^o(x) - \pi(x)\}b(x) = -\pi(x)L_b(x).$$

Finally, if $L_b(x) < 0$ and $U_b(x) \geq 0$ (so $\pi^*_{\mathbb{O}}(x) = 0$ and $\pi^*_{\mathbb{1}}(x) = 1$), then the oracle policy can be either 0 or 1, $\pi^o(x) \in \{0, 1\}$. Therefore,

$$\max_{e(x)\in[L(x),U(x)]} \{\pi^o(x) - \pi(x)\}b(x) = \max\{(1-\pi(x))U_b(x), -\pi(x)L_b(x)\} = U_b(x) - \pi(x)\{U_b(x) + L_b(x)\}.$$

Putting together the pieces, the worst-case regret relative to the oracle is

$$\begin{aligned} R_{\text{sup}}(\pi, \pi^o) = \mathbb{E}([\pi^*_{\mathbb{O}}(X) + \{1 - \pi^*_{\mathbb{O}}(X)\}\pi^*_{\mathbb{1}}(X)]U_b(X)) \\ - \mathbb{E}[\pi(X)\left\{\pi^*_{\mathbb{O}}(X)U_b(X) + (1 - \pi^*_{\mathbb{1}}(X))L_b(X) + (1 - \pi^*_{\mathbb{O}}(X))\pi^*_{\mathbb{1}}(X)(U_b(X) + L_b(X))\right\}], \end{aligned}$$

and the unconstrained minimizer is

$$\pi_o^* = \arg\min_{\pi} R_{\text{sup}}(\pi, \pi^o) = \begin{cases} \pi_{\mathbb{1}}^*(x), & \pi_{\mathbb{O}}^*(x) = 1, \\ \pi_{\mathbb{O}}^*(x), & \pi_{\mathbb{1}}^*(x) = 0, \\ \mathbb{1}\{U_b(x) \geq -L_b(x)\}, & \pi_{\mathbb{O}}^*(x) = 0, \pi_{\mathbb{1}}^*(x) = 1. \end{cases}$$

Now notice that $\pi_{\mathbb{O}}^*(x) = 1 \Leftrightarrow L_b(x) \geq 0 \Rightarrow U_b(x) \geq 0 \Leftrightarrow \pi_{\mathbb{1}}^*(x) = 1$, and $\pi_{\mathbb{1}}^*(x) = 0 \Leftrightarrow U_b(x) < 0 \Rightarrow L_b(x) < 0 \Leftrightarrow \pi_{\mathbb{O}}^*(x) = 0$, so we can simplify this to

$$\pi_o^* = \arg\min_{\pi} R_{\text{sup}}(\pi, \pi^o) = \begin{cases} 1, & \pi_{\mathbb{O}}^*(x) = 1, \\ 0, & \pi_{\mathbb{1}}^*(x) = 0, \\ \mathbb{1}\{U_b(x) \geq -L_b(x)\}, & \pi_{\mathbb{O}}^*(x) = 0, \pi_{\mathbb{1}}^*(x) = 1. \end{cases}$$

To complete the proof, we need to compute $L_b(x)$ and $U_b(x)$. First, we begin with the case where $u_g \geq u_l$. In this case,

$$L_b(x) = \{u_l + (u_g - u_l)\delta_\tau(x)\}\tau(x) - c = \{u_l + (u_g - u_l)\delta_\tau(x)\}m(1, x) - \{u_l + (u_g - u_l)\delta_\tau(x)\}m(0, x) - c,$$
$$U_b(x) = \{u_g - (u_g - u_l)\delta_+(x)\}m(1, x) - \{u_l + (u_g - u_l)\delta_+(x)\}m(0, x) + (u_g - u_l)\delta_+(x) - c.$$

This gives the form of the worst-case regret relative to $\pi^{\mathbb{1}}$ and $\pi^{\mathbb{O}}$. For the worst-case regret relative to the oracle, we collect terms to get

$$\begin{pmatrix} c_1^{\pi^o}(x) \\ c_0^{\pi^o}(x) \\ c^{\pi^o}(x) \end{pmatrix} = \begin{cases} (u_l + (u_g - u_l)\delta_\tau(x), -u_l - (u_g - u_l)\delta_\tau(x), -c), & \pi_{\mathbb{1}}^*(x) \\ (u_g - (u_g - u_l)\delta_+(x), -u_l - (u_g - u_l)\delta_+(x), (u_g - u_l)\delta_+(x) - c), & \pi_{\mathbb{O}}^*(x) \\ (u_l + u_g + (u_g - u_l)(\delta_\tau(x) - \delta_+(x)), -2u_l - (u_g - u_l)(\delta_\tau(x) + \delta_+(x)), (u_g - u_l)\delta_+(x) - 2c), & \pi_{\mathbb{O}}^*(x) \end{cases}$$

Now for the case where $u_g < u_l$, the lower and upper bounds switch:

$$L_b(x) = \{u_g - (u_g - u_l)\delta_+(x)\}m(1, x) - \{u_l + (u_g - u_l)\delta_+(x)\}m(0, x) + (u_g - u_l)\delta_+(x) - c,$$
$$U_b(x) = \{u_l + (u_g - u_l)\delta_\tau(x)\}\tau(x) - c = \{u_l + (u_g - u_l)\delta_\tau(x)\}m(1, x) - \{u_l + (u_g - u_l)\delta_\tau(x)\}m(0, x) - c.$$

For the worst-case regret relative to the oracle, we collect terms to get

$$\begin{pmatrix} c_1^{\pi^o}(x) \\ c_0^{\pi^o}(x) \\ c^{\pi^o}(x) \end{pmatrix} = \begin{cases} (u_g - (u_g - u_l)\delta_+(x), -u_l - (u_g - u_l)\delta_+(x), (u_g - u_l)\delta_+(x) - c), & \pi_{\mathbb{1}}^*(x) = \\ (u_l + (u_g - u_l)\delta_\tau(x), -u_l - (u_g - u_l)\delta_\tau(x), -c), & \pi_{\mathbb{O}}^*(x) = \\ (u_l + u_g + (u_g - u_l)(\delta_\tau(x) - \delta_+(x)), -2u_l - (u_g - u_l)(\delta_\tau(x) + \delta_+(x)), (u_g - u_l)\delta_+(x) - 2c), & \pi_{\mathbb{O}}^*(x) = \end{cases}$$

□

For the Proofs of Theorems 4.1 H.2 and 4.2, we prove the result for the case where $u_g \geq u_l$. The case where $u_g < u_l$ follows in the same way, with $\pi_{\mathbb{O}}$ taking the place for $\pi_{\mathbb{1}}$

Proof of Theorem 4.1. This follows directly from combining Lemmas J.1 and J.2 below via the union bound. □

Proof of Theorem 4.2. This follows directly from combining Lemmas J.1 and J.3 below via the union bound. □

Proof of Corollary 4.3. This follows by combining Theorem 4.1 and Lemma J.4 below. □

Proof of Theorem H.2. This follows from Lemmas J.4 and J.5 below. □

Proof of Corollary H.3. This follows by combining Theorem 4.2, Theorem H.2, and Lemma J.4 below. □

J.2 Auxiliary lemmas

Lemma J.1. Let $\hat{\pi}$ be a solution to Equation (10) with nuisance functions $\hat{m}$ and $\hat{d}$ fit on a separate sample, and let π^* be a solution to Equation (5). Under the strict overlap condition in Assumption 1, the excess worst-case regret between $\hat{\pi}$ and π^* is bounded by

$$R_{\text{sup}}(\hat{\pi}, \varpi) - R_{\text{sup}}(\pi^*, \varpi) \leq U \times \left\{ \frac{6+\eta}{\eta} \times \left(2\mathcal{R}_n(\Pi) + \frac{t}{\sqrt{n}} \right) + \sum_{w=0,1} \|\hat{\gamma}_w - \gamma_w\|_2 \, \|\hat{m}(w,\cdot) - m(w,\cdot)\|_2 \right\} + \sup_{\pi \in \Pi} \left| \tilde{R}(\pi, \hat{c}(\cdot), m(\cdot); \varpi) - \tilde{R}(\pi, c(\cdot), m(\cdot); \varpi) \right|,$$

with probability at least $1 - \exp\left(-\frac{t^2}{2}\right)$, where

$$\tilde{R}_{\text{sup}}(\pi, c(\cdot), m(\cdot); \varpi) = \frac{1}{n} \sum_{i=1}^{n} \pi(X_i) \left\{ c_1(X_i) m(1, X_i) + c_0(X_i) m(0, X_i) + c(X_i) \right\}.$$

and U is a constant depending on the utility values.

Proof of Lemma J.1. First, note that the excess regret can be decomposed into

$$\begin{aligned} R_{\text{sup}}(\hat{\pi}, \varpi) - R_{\text{sup}}(\pi^*, \varpi) &= R_{\text{sup}}(\hat{\pi}, \varpi) - \hat{R}_{\text{sup}}(\hat{\pi}, \varpi) + \underbrace{\hat{R}_{\text{sup}}(\hat{\pi}, \varpi) - \hat{R}_{\text{sup}}(\pi^*, \varpi)}_{\leq 0} + \hat{R}_{\text{sup}}(\pi^*, \varpi) - R_{\text{sup}}(\pi^*, \varpi) \\ &\leq 2 \sup_{\pi \in \Pi} |\hat{R}_{\text{sup}}(\pi, \varpi) - R_{\text{sup}}(\pi, \varpi)|, \end{aligned}$$

where we have used that $\hat{\pi}$ minimizes $\hat{R}_{\text{sup}}(\pi^*, \varpi)$.

We further decompose $\hat{R}_{\text{sup}}(\pi, \varpi) - R_{\text{sup}}(\pi, \varpi)$ into

$$\begin{aligned} \hat{R}_{\text{sup}}(\pi, \varpi) - R_{\text{sup}}(\pi, \varpi) &= \hat{R}_{\text{sup}}(\pi, \varpi) - \tilde{R}(\pi, \hat{c}(\cdot), m(\cdot); \varpi) && \text{(a)} \\ &\quad + \tilde{R}(\pi, c(\cdot), m(\cdot); \varpi) - R_{\text{sup}}(\pi, \varpi) && \text{(b)} \\ &\quad + \tilde{R}(\pi, \hat{c}(\cdot), m(\cdot); \varpi) - \tilde{R}(\pi, c(\cdot), m(\cdot); \varpi) \end{aligned}$$

We will now control terms (a) and (b), following closely the proof of Lemma 4 in Athey and Wager (2021). First note that we have the decompositions

$$\begin{aligned}\hat{\Gamma}_1(X,D,Y) - m(1,X) &= \hat{m}(1,X) - m(1,X) + \frac{D}{\hat{d}(X)}\left\{Y - \hat{m}(1,X)\right\} \\ &= \left\{\hat{m}(1,X) - m(1,X)\right\} \times \left(1 - \frac{D}{d(X)}\right) + \frac{D}{\hat{d}(X)}\left\{Y - m(1,X)\right\} \\ &\quad + \left(\frac{D}{\hat{d}(X)} - \frac{D}{d(X)}\right) \times \left\{m(1,X) - \hat{m}(1,X)\right\}\end{aligned}$$

and

$$\begin{aligned}\hat{\Gamma}_0(X,D,Y) - m(0,X) &= \hat{m}(0,X) - m(0,X) + \frac{1-D}{1-\hat{d}(X)}\left\{Y - \hat{m}(0,X)\right\} \\ &= \left\{\hat{m}(0,X) - m(0,X)\right\} \times \left(1 - \frac{1-D}{1-d(X)}\right) + \frac{1-D}{1-\hat{d}(X)}\left\{Y - m(0,X)\right\} \\ &\quad + \left(\frac{1-D}{1-\hat{d}(X)} - \frac{1-D}{1-d(X)}\right) \times \left\{m(0,X) - \hat{m}(0,X)\right\}.\end{aligned}$$

With this, we can compute the expectation of term (a):

$$\begin{aligned}\mathbb{E}[(a)] &= \mathbb{E}\left[\pi(X)\left(\hat{c}_1(X)\left\{\hat{\Gamma}_1(X,D,Y) - m(1,X)\right\} + \hat{c}_0(X)\left\{\hat{\Gamma}_0(X,D,Y) - m(0,X)\right\}\right)\right] \\ &= \mathbb{E}\left[\pi(X)\hat{c}_1(X)\left(\frac{D}{\hat{d}(X)} - \frac{D}{d(X)}\right) \times \left(m(1,X) - \hat{m}(1,X)\right)\right] \\ &\quad + \mathbb{E}\left[\pi(X)\hat{c}_0(X)\left(\frac{1-D}{1-\hat{d}(X)} - \frac{1-D}{1-d(X)}\right) \times \left(m(0,X) - \hat{m}(0,X)\right)\right],\end{aligned}$$

where we have used the fact that

$$\begin{aligned}\mathbb{E}\left[\pi(X)\hat{c}_1(X)\left(\hat{m}(1,X) - m(1,X)\right) \times \left(1 - \frac{D}{d(X)}\right)\right] &= 0, \\ \mathbb{E}\left[\pi(X)\hat{c}_1(X)\frac{D}{\hat{d}(X)}\left\{Y - m(1,X)\right\}\right] &= 0, \\ \mathbb{E}\left[\pi(X)\hat{c}_0(X)\left\{\hat{m}(0,X) - m(0,X)\right\} \times \left(1 - \frac{1-D}{1-d(X)}\right)\right] &= 0, \\ \mathbb{E}\left[\pi(X)\hat{c}_0(X)\frac{1-D}{1-\hat{d}(X)}\left\{Y - m(0,X)\right\}\right] &= 0,\end{aligned}$$

because $\hat{c}$, $\hat{m}$, and $\hat{d}$ come from a different sample.

The expectation of term (b) is

$$\mathbb{E}[(b)] = \mathbb{E}[\pi(X)\{c_1(X)m(1,X) + c_0(X)m(0,X) + c(X)\}] - R_{\text{sup}}(\pi, \varpi) = 0.$$

Now define a function $f : \mathcal{X} \to \mathbb{R}$ as

$$\begin{aligned} f_\pi(x,d,y) \;\equiv\; & \pi(x)\left[\hat{c}_1(x)\left\{\hat{\Gamma}_1(x,d,y) - m(1,x)\right\} + \hat{c}_0(x)\left\{\hat{\Gamma}_0(x,d,y) - m(0,x)\right\}\right] \\ & + \pi(x)\{c_1(x)m(1,x) + c_0(x)m(0,x) + c(x)\} \end{aligned}$$

and the function class $\mathcal{F} \equiv \{f_\pi \mid \pi \in \Pi\}$ as the set of all functions f as we vary π in Π.

With this notation, we can write the sum of terms (a) and (b) as

$$(a) + (b) = \frac{1}{n}\sum_{i=1}^{n} f_\pi(X_i, D_i, Y_i) - R_{\text{sup}}(\pi, \varpi),$$

and from above the expectation of $f_\pi(X_i, D_i, Y_i)$ is

$$\begin{aligned} \mathbb{E}[f_\pi(X,D,Y)] = R_{\text{sup}}(\pi,\varpi) &+ \mathbb{E}\left[\pi(X)\hat{c}_1(X)\left(\frac{D}{\hat{d}(X)} - \frac{D}{d(X)}\right) \times (m(1,X) - \hat{m}(1,X))\right] \\ &+ \mathbb{E}\left[\pi(X)\hat{c}_0(X)\left(\frac{1-D}{1-\hat{d}(X)} - \frac{1-D}{1-d(X)}\right) \times (m(0,X) - \hat{m}(0,X))\right]. \end{aligned}$$

Putting together the pieces, we can write

$$\begin{aligned} |(\text{a}) + (\text{b})| &= \left|\frac{1}{n}\sum_{i=1}^{n} f_\pi(X_i,D_i,Y_i) - R_{\text{sup}}(\pi,\varpi)\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^{n} f_\pi(X_i,D_i,Y_i) - \mathbb{E}[f_\pi(X,D,Y)] + \mathbb{E}[f_\pi(X,D,Y)] - R_{\text{sup}}(\pi,\varpi)\right| \\ &\le \left|\frac{1}{n}\sum_{i=1}^{n} f_\pi(X_i,D_i,Y_i) - \mathbb{E}[f_\pi(X,D,Y)]\right| \\ &\quad + \left|\mathbb{E}\left[\pi(X)\hat{c}_1(X)\left(\frac{D}{\hat{d}(X)} - \frac{D}{d(X)}\right) \times (m(1,X) - \hat{m}(1,X))\right]\right| \\ &\quad + \left|\mathbb{E}\left[\pi(X)\hat{c}_0(X)\left(\frac{1-D}{1-\hat{d}(X)} - \frac{1-D}{1-d(X)}\right) \times (m(0,X) - \hat{m}(0,X))\right]\right|. \end{aligned}$$

Now notice that for $\varpi \in \{\pi^{\mathbb{O}}, \pi^{\mathbb{1}}, \pi^{o}\}$, $|c_1(x)m(1,x) + c_0(x)m(0,x) + c(x)|$, $|c_1(x)|$, and $|c_0(x)|$ are bounded by some constant U depending on the utilities. From the decompositions

above, by the strict overlap condition in Assumption 1, and because $Y_i \in \{0, 1\}$,

$$\begin{aligned}
\left|\hat{\Gamma}_1(X_i, D_i, Y_i) - m(1, x)\right| &\leq \left|\{\hat{m}(1, X_i) - m(1, X_i)\} \times \left(1 - \frac{D_i}{d(X_i)}\right)\right| \\
&\quad + \left|\frac{D_i}{\hat{d}(X_i)} \times \{Y_i - m(1, X_i)\}\right| \\
&\quad + \left|\left(\frac{D_i}{d(X_i)} - \frac{D_i}{\hat{d}(X_i)}\right) \times \{\hat{m}(1, X_i) - m(1, X_i)\}\right| \\
&\leq \frac{1-\eta}{\eta}\|\hat{m} - m\|_\infty + \frac{1}{\eta} + \left\|\frac{1}{d} - \frac{1}{\hat{d}}\right\|_\infty \|\hat{m} - m\|_\infty \\
&\leq \frac{1-\eta}{\eta} + \frac{1}{\eta} + \frac{1}{\eta} - \frac{1}{1-\eta} \\
&\leq \frac{3}{\eta}.
\end{aligned}$$

Similarly,

$$\left|\hat{\Gamma}_0(X_i, D_i, Y_i) - m(0, x)\right| \leq \frac{1-\eta}{\eta}\|\hat{m} - m\|_\infty + \frac{1}{\eta} + \left\|\frac{1}{1-d} - \frac{1}{1-\hat{d}}\right\|_\infty \|\hat{m} - m\|_\infty \leq \frac{3}{\eta}.$$

This combines to give that for any x, d, y,

$$|f_\pi(x, d, y)| \leq U \times \frac{6+\eta}{\eta}.$$

This also shows that the Rademacher complexity of $\mathcal{F}$ is:

$$\mathcal{R}_n(\mathcal{F}) = 2U \times \frac{6+\eta}{\eta} \times \mathcal{R}_n(\Pi).$$

So by Wainwright (2019) Theorem 4.2, for any $n \geq 1$ and $t \geq 0$,

$$\sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right| \leq 2U \times \frac{6+\eta}{\eta} \times \left(2\mathcal{R}_n(\Pi) + \frac{t}{\sqrt{n}}\right),$$

with probability at least $1 - \exp\left(-\frac{t^2}{2}\right)$.

Finally, notice that by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left|\mathbb{E}\left[\pi(X)\hat{c}_1(X)\left(\frac{D}{\hat{d}(X)} - \frac{D}{d(X)}\right) \times (m(1, X) - \hat{m}(1, X))\right]\right| \\
&\leq U\sqrt{\mathbb{E}\left[\left(\frac{D}{\hat{d}(X)} - \frac{D}{d(X)}\right)^2\right]\mathbb{E}\left[(m(1, X) - \hat{m}(1, X))^2\right]},
\end{aligned}$$

and

$$\left|\mathbb{E}\left[\pi(X)\hat{c}_0(X)\left(\frac{1-D}{1-\hat{d}(X)}-\frac{1-D}{1-d(X)}\right)\times(m(0,X)-\hat{m}(0,X))\right]\right|$$
$$\leq U\sqrt{\mathbb{E}\left[\left(\frac{1-D}{1-\hat{d}(X)}-\frac{1-D}{1-d(X)}\right)^2\right]\mathbb{E}\left[(m(0,X)-\hat{m}(0,X))^2\right]}.$$

Combining these two bounds gives the result. □

Lemma J.2. For $u_g \geq u_l$,

$$\sup_{\pi\in\Pi}\left|\tilde{R}_{\text{sup}}(\pi,\hat{c},m;\pi^*_{\mathbb{1}})-\tilde{R}_{\text{sup}}(\pi,c,m,\pi^*_{\mathbb{1}})\right|\leq(u_g-u_l)\times\left(R_+(\hat{\delta}_+)+\frac{t}{2\sqrt{n}}\right),$$

with probability at least $1-e^{-\frac{t^2}{2}}$.

Proof of Lemma J.2. First we have the bound,

$$\begin{aligned}\tilde{R}_{\text{sup}}(\pi,\hat{c},m;\pi^*_{\mathbb{1}})-\tilde{R}_{\text{sup}}(\pi,c,m,\pi^*_{\mathbb{1}}) &= \frac{u_g-u_l}{n}\sum_{i=1}^n\pi(X_i)\left\{\hat{\delta}_+(X_i)-\delta_+(X_i)\right\}\{m(1,X_i)+m(0,X_i)-1\}\\ &\leq \frac{u_g-u_l}{n}\sum_{i=1}^n\mathbb{1}\left\{\hat{\delta}_+(X_i)\neq\delta_+(X_i)\right\}|m(1,X_i)+m(0,X_i)-1|.\end{aligned}$$

Now note that

$$\mathbb{E}\left[\mathbb{1}\left\{\hat{\delta}_+(X_i)\neq\delta_+(X_i)\right\}|m(1,X_i)+m(0,X_i)-1|\right]=R_+(\hat{\delta}_+)$$

For each i, since $\mathbb{1}\left\{\hat{\delta}_+(X_i)\neq\delta_+(X_i)\right\}|m(1,X_i)+m(0,X_i)-1|$ is bounded between 0 and 1, it is sub-Gaussian with scale parameter 1. Furthermore, they are independent across $i=1,\ldots,n$, so by the Hoeffding bound,

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n\mathbb{1}\left\{\hat{\delta}_+(X_i)\neq\delta_+(X_i)\right\}|m(1,X_i)+m(0,X_i)-1|\leq R_+(\hat{\delta}_+)+\frac{t}{\sqrt{n}}\right)\geq 1-\exp\left(-2t^2\right).$$

Combining this with the deterministic bound above gives the result.

□

Lemma J.3. For $u_g \geq u_l$,

$$\begin{aligned}&\sup_{\pi\in\Pi}\left|\tilde{R}_{\text{sup}}(\pi,\hat{c},m;\pi^*_o)-\tilde{R}_{\text{sup}}(\pi,c,m,\pi^*_o)\right|\\ \leq\ & 2\times\left\{R_{\text{sup}}(\hat{\pi}_{\mathbb{1}},\pi^{\mathbb{1}})-R_{\text{sup}}(\pi^*_{\mathbb{1}},\pi^{\mathbb{1}})\right\}+2\times\left\{R_{\text{sup}}(\hat{\pi}_{\mathbb{O}},\pi^{\mathbb{O}})-R_{\text{sup}}(\pi^*_{\mathbb{O}},\pi^{\mathbb{O}})\right\}\\ &+(u_g-u_l)\times\left(R_+(\hat{\delta}_+)+R_\tau(\hat{\delta}_\tau)+\frac{t}{2\sqrt{n}}\right),\end{aligned}$$

with probability at least $1 - 2e^{-\frac{t^2}{2}}$.

Proof of Lemma J.3. Define

$$
\begin{aligned}
\check{L}_b(x) &= \{u_l + (u_g - u_l)\hat{\delta}_\tau(x)\}m(1,x) - \{u_l + (u_g - u_l)\hat{\delta}_\tau(x)\}m(0,x) - c,\\
\check{U}_b(x) &= \{u_g - (u_g - u_l)\hat{\delta}_+(x)\}m(1,x) - \{u_l + (u_g - u_l)\hat{\delta}_+(x)\}m(0,x) + (u_g - u_l)\hat{\delta}_+(x) - c,\\
Q(x) &= \pi^*_{\mathbb{O}}(x)U_b(x) + (1 - \pi^*_{\mathbb{1}}(x))L_b(x) + (1 - \pi^*_{\mathbb{O}}(x))\pi^*_{\mathbb{1}}(x)(U_b(x) + L_b(x)),\\
\tilde{Q}(x) &= \hat{\pi}_{\mathbb{O}}(x)U_b(x) + (1 - \hat{\pi}_{\mathbb{1}}(x))L_b(x) + (1 - \hat{\pi}_{\mathbb{O}}(x))\hat{\pi}_{\mathbb{1}}(x)(U_b(x) + L_b(x)),\\
\check{Q}(x) &= \hat{\pi}_{\mathbb{O}}(x)\check{U}_b(x) + (1 - \hat{\pi}_{\mathbb{1}}(x))\check{L}_b(x) + (1 - \hat{\pi}_{\mathbb{O}}(x))\hat{\pi}_{\mathbb{1}}(x)(\check{U}_b(x) + \check{L}_b(x)).
\end{aligned}
$$

With these definitions, we can write

$$
\begin{aligned}
\left|\tilde{R}_{\text{sup}}(\pi, \hat{c}, m; \pi^*_o) - \tilde{R}_{\text{sup}}(\pi, c, m, \pi^*_o)\right| &= \left|\frac{1}{n}\sum_{i=1}^n \pi(X)\{Q(X_i) - \check{Q}(X_i)\}\right|\\
&= \left|\frac{1}{n}\sum_{i=1}^n \pi(X)\{Q(X_i) - \tilde{Q}(X_i)\} + \frac{1}{n}\sum_{i=1}^n \pi(X)\{\tilde{Q}(X_i) - \check{Q}(X_i)\}\right|\\
&\leq \frac{1}{n}\sum_{i=1}^n \left|Q(X_i) - \tilde{Q}(X_i)\right| + \frac{1}{n}\sum_{i=1}^n \left|\tilde{Q}(X_i) - \check{Q}(X_i)\right|.
\end{aligned}
$$

Working with the first term:

$$
\begin{aligned}
Q(x) - \tilde{Q}(x) &= (\hat{\pi}_{\mathbb{1}}(x) - \pi^*_{\mathbb{1}}(x))U_b(x) - (\hat{\pi}_{\mathbb{1}}(x)\hat{\pi}_{\mathbb{O}}(x) - \pi^*_{\mathbb{1}}(x)\pi^*_{\mathbb{O}}(x))(U_b(x) + L_b(x)) + (\hat{\pi}_{\mathbb{O}}(x) - \pi^*_{\mathbb{O}}(x))U_b(\\
&= (\hat{\pi}_{\mathbb{O}}(x) - \pi^*_{\mathbb{O}}(x)) \times (-L_b(x)\pi^*_{\mathbb{1}}(x) + (1 - \pi^*_{\mathbb{1}}(x))U_b(x)) \qquad (*)\\
&\quad + (\hat{\pi}_{\mathbb{1}}(x) - \pi^*_{\mathbb{1}}(x)) \times (-L_b(x)\hat{\pi}_{\mathbb{O}}(x) + (1 - \hat{\pi}_{\mathbb{O}}(x))U_b(x)) \qquad (**)
\end{aligned}
$$

Notice that $\pi^*_{\mathbb{1}}(x) = 0 \Leftrightarrow U_b(x) \leq 0$, since $L_b(x) \leq U_b(x)$, this implies that when $\pi^*_{\mathbb{1}}(x) = 0$, $|U_b(x)| \leq |L_b(x)|$. Therefore,

$$
|(*)| \leq \mathbb{1}\{\hat{\pi}_{\mathbb{O}}(x) \neq \pi^*_{\mathbb{O}}(x)\}|L_b(x)|.
$$

Similarly, if $\pi^*_{\mathbb{O}}(x) = 1$, then $0 \leq L_b(x) \leq U_b(x)$, so $|L_b(x)| \leq |U_b(x)|$. So,

$$
\begin{aligned}
|(**)| &\leq \mathbb{1}\{\hat{\pi}_{\mathbb{1}}(x) \neq \pi^*_{\mathbb{1}}(x)\}\mathbb{1}\{\hat{\pi}_{\mathbb{O}}(x) = \hat{\pi}_{\mathbb{O}}(x)\}| - L_b(x)\pi^*_{\mathbb{O}}(x) + (1 - \pi^*_{\mathbb{O}}(x))U_b(x)|\\
&\quad + \mathbb{1}\{\hat{\pi}_{\mathbb{O}}(x) \neq \pi^*_{\mathbb{O}}(x)\}\mathbb{1}\{\hat{\pi}_{\mathbb{1}}(x) \neq \pi^*_{\mathbb{1}}(x)\}| - \hat{\pi}_{\mathbb{O}}(x)L_b(x) + (1 - \hat{\pi}_{\mathbb{O}}(x))U_b(x)|\\
&\leq \mathbb{1}\{\hat{\pi}_{\mathbb{1}}(x) \neq \pi^*_{\mathbb{1}}(x)\}\mathbb{1}\{\hat{\pi}_{\mathbb{O}}(x) = \hat{\pi}_{\mathbb{O}}(x)\}|U_b(x)|\\
&\quad + \mathbb{1}\{\hat{\pi}_{\mathbb{O}}(x) \neq \pi^*_{\mathbb{O}}(x)\}\mathbb{1}\{\hat{\pi}_{\mathbb{1}}(x) \neq \pi^*_{\mathbb{1}}(x)\}|L_b(x)| + \mathbb{1}\{\hat{\pi}_{\mathbb{O}}(x) \neq \pi^*_{\mathbb{O}}(x)\}\mathbb{1}\{\hat{\pi}_{\mathbb{1}}(x) \neq \pi^*_{\mathbb{1}}(x)\}|U_b(x)|\\
&\leq \mathbb{1}\{\hat{\pi}_{\mathbb{1}}(x) \neq \pi^*_{\mathbb{1}}(x)\}|U_b(x)| + \mathbb{1}\{\hat{\pi}_{\mathbb{O}}(x) \neq \pi^*_{\mathbb{O}}(x)\}|L_b(x)| + \mathbb{1}\{\hat{\pi}_{\mathbb{1}}(x) \neq \pi^*_{\mathbb{1}}(x)\}|U_b(x)|\\
&\leq 2\mathbb{1}\{\hat{\pi}_{\mathbb{1}}(x) \neq \pi^*_{\mathbb{1}}(x)\}|U_b(x)| + \mathbb{1}\{\hat{\pi}_{\mathbb{O}}(x) \neq \pi^*_{\mathbb{O}}(x)\}|L_b(x)|.
\end{aligned}
$$

Putting together the pieces, we get that

$$|Q(x) - \tilde{Q}(x)| \leq 2\mathbb{1}\{\hat{\pi}_{\mathbb{O}}(x) \neq \pi^*_{\mathbb{O}}(x)\}|L_b(x)| + 2\mathbb{1}\{\hat{\pi}_{\mathbb{1}}(x) \neq \pi^*_{\mathbb{1}}(x)\}|U_b(x)|.$$

So the expectation is bounded by two regret terms:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|Q(X_i) - \tilde{Q}(X_i)\right|\right] &\leq 2\mathbb{E}\left[\mathbb{1}\{\hat{\pi}_{\mathbb{O}}(X) \neq \pi^*_{\mathbb{O}}(X)\}|L_b(X)|\right] + 2\mathbb{E}\left[\mathbb{1}\{\hat{\pi}_{\mathbb{1}}(X) \neq \pi^*_{\mathbb{1}}(X)\}|U_b(X)|\right] \\ &= 2 \times \{R_{\text{sup}}(\hat{\pi}_{\mathbb{1}}, \pi^{\mathbb{1}}) - R_{\text{sup}}(\pi^*_{\mathbb{1}}, \pi^{\mathbb{1}})\} + 2 \times \{R_{\text{sup}}(\hat{\pi}_{\mathbb{O}}, \pi^{\mathbb{O}}) - R_{\text{sup}}(\pi^*_{\mathbb{O}}, \pi^{\mathbb{O}})\}.\end{aligned}$$

Next, $\left|Q(X_i) - \tilde{Q}(X_i)\right|$ is bounded between 0 and $u_g - u_l$, so by the Hoeffding bound it concentrates around its expectation:

$$\begin{aligned}&\Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left|Q(X_i) - \tilde{Q}(X_i)\right| \leq 2\{R_{\text{sup}}(\hat{\pi}_{\mathbb{1}}, \pi^{\mathbb{1}}) - R_{\text{sup}}(\pi^*_{\mathbb{1}}, \pi^{\mathbb{1}})\} + 2\{R_{\text{sup}}(\hat{\pi}_{\mathbb{O}}, \pi^{\mathbb{O}}) - R_{\text{sup}}(\pi^*_{\mathbb{O}}, \pi^{\mathbb{O}})\} + \frac{t}{\sqrt{}}\right. \\ \geq\ & 1 - \exp\left(-\frac{2t^2}{(u_g - u_l)^2}\right).\end{aligned}$$

Now for the second term:

$$\begin{aligned}|\tilde{Q}(x) - \check{Q}(x)| &= \left|(L_b(x) - \check{L}_b(x))(1 - \hat{\pi}_{\mathbb{1}} + (1 - \hat{\pi}_{\mathbb{O}})\hat{\pi}_{\mathbb{1}}) + (U_b(x) - \check{U}_b(x))(\hat{\pi}_{\mathbb{O}} + (1 - \hat{\pi}_{\mathbb{O}})\hat{\pi}_{\mathbb{1}})\right| \\ &\leq |L_b(x) - \check{L}_b(x)| + |U_b(x) - \check{U}_b(x)|.\end{aligned}$$

To re-write this, notice that

$$\begin{aligned}|L_b(x) - \check{L}_b(x)| &= (u_g - u_l)\mathbb{1}\{\hat{\delta}_\tau(x) \neq \delta_\tau(x)\}|m(1,x) - m(0,x)|, \\ |U_b(x) - \check{U}_b(x)| &= (u_g - u_l)\mathbb{1}\{\hat{\delta}_+(x) \neq \delta_+(x)\}|m(1,x) + m(0,x) - 1|.\end{aligned}$$

So,

$$\frac{|\tilde{Q}(x) - \check{Q}(x)|}{u_g - u_l} \leq \mathbb{1}\{\hat{\delta}_\tau(x) \neq \delta_\tau(x)\}|m(1,x) - m(0,x)| + \mathbb{1}\{\hat{\delta}_+(x) \neq \delta_+(x)\}|m(1,x) + m(0,x) - 1|.$$

Taking the expectation, we see that it is bounded by:

$$\begin{aligned}\frac{1}{u_g - u_l}\frac{1}{n}\sum_{i=1}^{n}\left|\tilde{Q}(X_i) - \check{Q}(X_i)\right| &\leq \mathbb{E}\left[\mathbb{1}\{\hat{\delta}_\tau(x) \neq \delta_\tau(X)\}|m(1,X) - m(0,X)|\right] \\ &\quad + \mathbb{E}\left[\mathbb{1}\{\hat{\delta}_+(X) \neq \delta_+(x)\}|m(1,X) + m(0,X) - 1|\right] \\ &= R_+(\hat{\delta}_+) + R_\tau(\hat{\delta}_\tau).\end{aligned}$$

Again noting that $|\tilde{Q}(X_i) - \check{Q}(X_i)|$ is bounded between 0 and $u_g - u_l$, and applying the Hoeffding inequality gives

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left|\tilde{Q}(X_i)-\check{Q}(X_i)\right| \le (u_g-u_l)\times\left(R_+(\hat{\delta}_+)+R_\tau(\hat{\delta}_\tau)+\frac{t}{\sqrt{n}}\right)\right) \ge 1-\exp\left(-2t^2\right).$$

Combining these two bounds via the union bound gives the result.

□

Lemma J.4. Let $\hat{\delta}_+(x) = \mathbb{1}\{\hat{m}(1,x)+\hat{m}(0,x)-1\ge 0\}$ and $\hat{\delta}_\tau(x) = \mathbb{1}\{\hat{m}(1,x)-\hat{m}(0,x)\}$. Under Assumption 2,

$$\begin{aligned} R_+(\hat{\delta}_+) &\le 2^{1+\alpha}C\|\hat{m}-m\|_\infty^{1+\alpha}, \\ \Pr(\hat{\delta}_+(X)\neq\delta_+(X)) &\le 2^{\alpha}C\|\hat{m}-m\|_\infty^{\alpha}. \end{aligned}$$

Under Assumption H.1(b),

$$\begin{aligned} R_\tau(\hat{\delta}_\tau) &\le 2^{1+\alpha}C\|\hat{m}-m\|_\infty^{1+\alpha}, \\ \Pr(\hat{\delta}_\tau(X)\neq\delta_\tau(X)) &\le 2^{\alpha}C\|\hat{m}-m\|_\infty^{\alpha}. \end{aligned}$$

Proof of Lemma J.4. This Lemma directly follows Lemma 5.1 in Audibert and Tsybakov (2007). Note that if $\hat{\delta}_+(x)\neq\delta_+(x)$, then the error is greater than the margin, i.e.,

$$|\hat{m}(1,x)-m(1,x)+\hat{m}(0,x)-m(0,x)| \ge |m(1,X)+m(0,X)-1|$$

So,

$$\begin{aligned} \Pr(\hat{\delta}_+(X)\neq\delta_+(X)) &\le \Pr(|\hat{m}(1,X)-m(1,X)+\hat{m}(0,X)-m(0,X)| \ge |m(1,X)+m(0,X)-1|) \\ &\le C(\|\hat{m}(1,\cdot)-m(1,\cdot)\|_\infty+\|\hat{m}(0,\cdot)-m(0,\cdot)\|_\infty)^{\alpha}. \end{aligned}$$

By a similar argument,

$$\begin{aligned} R_+(\hat{\delta}_+)-R_+(\delta_+) &= \mathbb{E}\left[\mathbb{1}\left\{\hat{\delta}_+(X)\neq\delta_+(X)\right\}|m(1,X)+m(0,X)-1|\right] \\ &\le \mathbb{E}\left[\mathbb{1}\left\{|\hat{m}(1,X)-m(1,X)+\hat{m}(0,X)-m(0,X)| \ge |m(1,X)+m(0,X)-1|\right\}\right. \\ &\qquad \left.\times\,|m(1,X)+m(0,X)-1|\right] \\ &\le \mathbb{E}\left[\mathbb{1}\left\{|\hat{m}(1,X)-m(1,X)+\hat{m}(0,X)-m(0,X)| \ge |m(1,X)+m(0,X)-1|\right\}\right. \\ &\qquad \left.\times\,|m(1,X)-\hat{m}(1,X)+m(0,X)-\hat{m}(0,X)|\right] \\ &\le (\|\hat{m}(1,\cdot)-m(1,\cdot)\|_\infty+\|\hat{m}(0,\cdot)-m(0,\cdot)\|_\infty) \\ &\qquad \times\Pr(|\hat{m}(1,X)-m(1,X)+\hat{m}(0,X)-m(0,X)| \ge |m(1,X)+m(0,X)-1|) \\ &\le C(\|\hat{m}(1,\cdot)-m(1,\cdot)\|_\infty+\|\hat{m}(0,\cdot)-m(0,\cdot)\|_\infty)^{1+\alpha}. \end{aligned}$$

Similarly, if $\hat{\delta}_\tau(x) \neq \delta_\tau(x)$, then

$$|m(1,x) - \hat{m}(1,x) - m(0,x) + \hat{m}(0,x)| \geq |m(1,x) - m(0,x)|.$$

By the same argument as above,

$$\Pr(\hat{\delta}_\tau(X) \neq \delta_\tau(X)) \leq C(\|\hat{m}(1,\cdot) - m(1,\cdot)\|_\infty + \|\hat{m}(1,\cdot) - m(1,\cdot)\|_\infty)^\alpha,$$

and

$$\begin{aligned}
R_\tau(\hat{\delta}_\tau) - R_\tau(\delta_\tau) &= \mathbb{E}\left[\mathbb{1}\left\{\hat{\delta}_\tau(X_i) \neq \delta_\tau(X_i)\right\}|m(1,X_i) - m(0,X_i)|\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\left\{|m(1,X) - \hat{m}(1,X) - m(0,X) + \hat{m}(0,X)| \geq |m(1,X) - m(0,X)|\right\}\right. \\
&\qquad \left. \times\ |m(1,X) - m(0,X)|\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\left\{|m(1,X) - \hat{m}(1,X) - m(0,X) + \hat{m}(0,X)| \geq |m(1,X) - m(0,X)|\right\}\right. \\
&\qquad \left. \times\ |m(1,X) - \hat{m}(1,X) - m(0,X) + \hat{m}(0,X)|\right] \\
&\leq (\|\hat{m}(1,\cdot) - m(1,\cdot)\|_\infty + \|\hat{m}(0,\cdot) - m(0,\cdot)\|_\infty) \\
&\qquad \times \Pr(|m(1,X) - \hat{m}(1,X) - m(0,X) + \hat{m}(0,X)| \geq |m(1,X) - m(0,X)|) \\
&\leq C(\|\hat{m}(1,\cdot) - m(1,\cdot)\|_\infty + \|\hat{m}(0,\cdot) - m(0,\cdot)\|_\infty)^{1+\alpha}.
\end{aligned}$$

□

Lemma J.5. Let $u_g \geq u_l$. Define

$$\begin{aligned}
\widehat{L}_b(x) &= \{u_l + \hat{\delta}_\tau(x)(u_g - u_l)\}\{\hat{m}(1,x) - \hat{m}(0,x)\} - c, \\
\widehat{U}_b(x) &= \{u_g - (u_g - u_l)\hat{\delta}_+(x)\}\hat{m}(1,x) - \{u_l + (u_g - u_l)\hat{\delta}_+(x)\}\hat{m}(0,x) + (u_g - u_l)\hat{\delta}_+(x) - c.
\end{aligned}$$

and let $\hat{\pi}^{\text{plug}}_{\mathbb{0}}(x) = \mathbb{1}\{\widehat{L}_b(x) \geq 0\}$ and $\hat{\pi}^{\text{plug}}_{\mathbb{1}}(x) = \mathbb{1}\{\widehat{U}_b(x) \geq 0\}$ be the plug-in estimates of the minimax optimal policies relative to never or always treating. Under Assumption H.1(d), the excess worst case regret for $\hat{\pi}^{\text{plug}}_{\mathbb{0}}$ relative to $\pi^*_{\mathbb{0}}$ is

$$R_{\text{sup}}(\hat{\pi}^{\text{plug}}_{\mathbb{0}}, \pi^{\mathbb{0}}) - R_{\text{sup}}(\pi^*_{\mathbb{0}}, \pi^{\mathbb{0}}) \leq u_g^\alpha C(2\|m - \hat{m}\|_\infty)^{1+\alpha} + 2u_g C\|m - \hat{m}\|_\infty \Pr\left(\hat{\delta}_\tau(X) \neq \delta_\tau(X)\right) + (u_g - u_l)R_\tau(\hat{\delta}_\tau$$

Under Assumption H.1(c), the excess worst case regret for $\hat{\pi}^{\text{plug}}_{\mathbb{1}}$ relative to $\pi^*_{\mathbb{1}}$ is

$$R_{\text{sup}}(\hat{\pi}^{\text{plug}}_{\mathbb{1}}, \pi^{\mathbb{1}}) - R_{\text{sup}}(\pi^*_{\mathbb{0}}, \pi^{\mathbb{0}}) \leq u_g^\alpha C(2\|m - \hat{m}\|_\infty)^{1+\alpha} + 2u_g C\|m - \hat{m}\|_\infty \Pr\left(\hat{\delta}_+(X) \neq \delta_+(X)\right) + (u_g - u_l)R_+(\hat{\delta}_-$$

Proof of Lemma J.5. First, as in the proof of Lemma J.4, note that $\hat{\pi}^{\text{plug}}_{\mathbb{0}}(x) \neq \pi^*_{\mathbb{0}}(x)$ implies that

$|L_b(x) - \widehat{L}_b(x)| \geq |L_b(x)|$. Now, if $\hat{\delta}_\tau(x) = \delta_\tau(x)$, then

$$\begin{aligned}|L_b(x) - \widehat{L}_b(x)| &= |((1-\delta_\tau(x))u_l + \delta_\tau(x)u_g)(m(1,x) - \hat{m}(1,x) - m(0,x) + \hat{m}(0,x))| \\ &\leq u_g |m(1,x) - \hat{m}(1,x) - m(0,x) + \hat{m}(0,x)|,\end{aligned}$$

because $|(1-\delta_\tau(x))u_l + \delta_\tau(x)u_g| = |u_l + (u_g - u_l)\delta_\tau(X)| \leq \max\{u_g, u_l\} \leq u_g$ in the case where $u_g \geq u_l$. If $\hat{\delta}_\tau(x) \neq \delta_\tau(x)$ and $\delta_\tau(x) = 1$, we have that

$$\begin{aligned}|L_b(x) - \widehat{L}_b(x)| &= |u_l(m(1,x) - \hat{m}(1,x) - m(0,x) + \hat{m}(0,x)) + (u_g - u_l)(m(1,x) - m(0,x))| \\ &\leq u_l |m(1,x) - \hat{m}(1,x) - m(0,x) + \hat{m}(0,x)| + (u_g - u_l)|m(1,x) - m(0,x)| \\ &\leq u_g |m(1,x) - \hat{m}(1,x) - m(0,x) + \hat{m}(0,x)| + (u_g - u_l)|m(1,x) - m(0,x)|.\end{aligned}$$

Similarly, if $\hat{\delta}_\tau(x) \neq \delta_\tau(x)$ and $\delta_\tau(x) = 0$,

$$\begin{aligned}|L_b(x) - \widehat{L}_b(x)| &= |u_g(m(1,x) - \hat{m}(1,x) - m(0,x) + \hat{m}(0,x)) - (u_g - u_l)(m(1,x) - m(0,x))| \\ &\leq u_g |m(1,x) - \hat{m}(1,x) - m(0,x) + \hat{m}(0,x)| + (u_g - u_l)|m(1,x) - m(0,x)|.\end{aligned}$$

Putting together the pieces, we get that

$$\begin{aligned}R_{\text{sup}}(\hat{\pi}^{\text{plug}}_{\mathbb{O}}, \pi^{\mathbb{O}}) - R_{\text{sup}}(\pi^*_{\mathbb{O}}, \pi^{\mathbb{O}}) &= \mathbb{E}\left[\mathbb{1}\{\hat{\pi}^{\text{plug}}_{\mathbb{O}} \neq \pi^*_{\mathbb{O}}\}|L_b(x)|\right] \\ &\leq \mathbb{E}\left[\mathbb{1}\{|L_b(X) - \widehat{L}_b(X)| \geq |L_b(X)|\}|L(X)|\right] \\ &\leq \mathbb{E}\left[\mathbb{1}\{|L_b(X) - \widehat{L}_b(X)| \geq |L_b(X)|\}|L_b(X) - \widehat{L}_b(X)|\right] \\ &= \mathbb{E}\left[\mathbb{1}\{|L_b(X) - \widehat{L}_b(X)| \geq |L_b(X)|\}|L_b(X) - \widehat{L}_b(X)|\mathbb{1}\{\hat{\delta}_\tau(X) = \delta_\tau(X)\}\right] \\ &\qquad (*) \\ &\quad + \mathbb{E}\left[\mathbb{1}\{|L_b(X) - \widehat{L}_b(X)| \geq |L_b(X)|\}|L_b(X) - \widehat{L}_b(X)|\mathbb{1}\{\hat{\delta}_\tau(X) \neq \delta_\tau(X)\right. \\ &\qquad (**)\end{aligned}$$

By Hölder's inequality and the margin condition (Assumption H.1(d)), the first term is

$$\begin{aligned}(*) &\leq \mathbb{E}\left[\mathbb{1}\{|((1-\delta_\tau(x))u_l + \delta_\tau(X)u_g)(m(1,x) - \hat{m}(1,x) - m(0,x) + \hat{m}(0,x))| \geq |L(X)|\}\right. \\ &\qquad \left.\times\ |m(1,X) - \hat{m}(1,X) - m(0,X) + \hat{m}(0,X)|\right] \\ &\leq \mathbb{E}\left[\mathbb{1}\{u_g|m(1,x) - \hat{m}(1,x) - m(0,x) + \hat{m}(0,x)| \geq |L(X)|\}\right. \\ &\qquad \left.\times\ |m(1,X) - \hat{m}(1,X) - m(0,X) + \hat{m}(0,X)|\right] \\ &\leq \mathbb{E}\left[\mathbb{1}\{u_g|m(1,x) - \hat{m}(1,x) - m(0,x) + \hat{m}(0,x)| \geq |L(X)|\}\right] \times 2\|m - \hat{m}\|_\infty \\ &\leq C u_g^\alpha (2\|m - \hat{m}\|_\infty)^{1+\alpha}.\end{aligned}$$

Similarly, we can bound the second term as

$$\begin{aligned}(**) &\leq \mathbb{E}\left[|L_b(X) - \widehat{L}_b(X)|\mathbb{1}\{\hat{\delta}_\tau(X) \neq \delta_\tau(X)\}\right] \\ &\leq \mathbb{E}\left[u_g|m(1,X) - \hat{m}(1,X) - m(0,X) + \hat{m}(0,X)|\mathbb{1}\{\hat{\delta}_\tau(X) \neq \delta_\tau(X)\}\right] \\ &\quad + (u_g - u_l)\mathbb{E}\left[|m(1,X) - m(0,X)|\mathbb{1}\{\hat{\delta}_\tau(X) \neq \delta_\tau(X)\}\right] \\ &\leq u_g 2C\|m - \hat{m}\|_\infty \Pr\left(\hat{\delta}_\tau(X) \neq \delta_\tau(X)\right) + (u_g - u_l)R_\tau(\hat{\delta}_\tau).\end{aligned}$$

Combining these two terms gives the first result.

Now, also note that $\hat{\pi}_{\mathbb{1}}^{\text{plug}}(x) \neq \pi_{\mathbb{1}}^*(x)$ implies that $|U_b(x) - \widehat{U}_b(x)| \geq |U_b(x)|$. We again break this error term into cases depending on $\hat{\delta}_+(x)$ and $\delta_+(x)$. First, if $\hat{\delta}_+(x) = \delta_+(x)$, then

$$\begin{aligned}|U_b(x) - \widehat{U}_b(x)| &= \begin{cases} |u_g(m(1,x) - \hat{m}(1,x)) - u_l(m(0,x) - \hat{m}(0,x))|, & \delta_+(x) = 0 \\ |u_l(m(1,x) - \hat{m}(1,x)) - u_g(m(0,x) - \hat{m}(0,x))|, & \delta_+(x) = 1 \end{cases} \\ &\leq u_g|m(1,x) - \hat{m}(1,x)| + u_g|m(0,x) - \hat{m}(0,x)|.\end{aligned}$$

If $\hat{\delta}_+(x) \neq \delta_+(x)$

$$\begin{aligned}|U_b(x) - \widehat{U}_b(x)| &= \begin{cases} |u_g(m(1,x) - \hat{m}(1,x)) - u_l(m(0,x) - \hat{m}(0,x)) + (u_g - u_l)(m(1,x) + m(0,x) - 1)|, & \\ |u_l(m(1,x) - \hat{m}(1,x)) - u_g(m(0,x) - \hat{m}(0,x))| - (u_g - u_l)(m(1,x) + m(0,x) - 1), & \end{cases} \\ &\leq u_g|m(1,x) - \hat{m}(1,x)| + u_g|m(0,x) - \hat{m}(0,x)| + (u_g - u_l)|m(1,x) + m(0,x) - 1|.\end{aligned}$$

Mirroring the decomposition above, we have that

$$
\begin{aligned}
R_{\text{sup}}(\hat{\pi}_{\mathbb{1}}^{\text{plug}}, \pi^{\mathbb{1}}) - R_{\text{sup}}(\pi_{\mathbb{1}}^{*}, \pi^{\mathbb{1}}) &= \mathbb{E}\left[\mathbb{1}\{\hat{\pi}_{\mathbb{1}}^{\text{plug}} \neq \pi_{\mathbb{1}}^{*}\}|U_b(x)|\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\{|U_b(X) - \widehat{U}_b(X)| \geq |U_b(X)|\}|U_b(X)|\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\{|U_b(X) - \widehat{U}_b(X)| \geq |U_b(X)|\}|U_b(X) - \widehat{U}_b(X)|\right] \\
&= \mathbb{E}\left[\mathbb{1}\{|U_b(X) - \widehat{U}_b(X)| \geq |U_b(X)|\}|U_b(X) - \widehat{U}_b(X)|\mathbb{1}\{\hat{\delta}_+(X) = \delta_+(X)\}\right] \\
&\quad + \mathbb{E}\left[\mathbb{1}\{|U_b(X) - \widehat{U}_b(X)| \geq |U_b(X)|\}|U_b(X) - \widehat{U}_b(X)|\mathbb{1}\{\hat{\delta}_+(X) \neq \delta_+(X)\}\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\{u_g|m(1,X) - \hat{m}(1,X)| + u_g|m(0,X) - \hat{m}(0,X)| \geq |U_b(X)|\}\right. \\
&\qquad \left. \times (u_g|m(1,X) - \hat{m}(1,X)| + u_g|m(0,X) - \hat{m}(0,X)|)\right] \\
&\quad + \mathbb{E}\left[u_g|m(1,x) - \hat{m}(1,x)|\mathbb{1}\{\hat{\delta}_+(X) \neq \delta_+(X)\}\right] \\
&\quad + \mathbb{E}\left[u_g|m(0,x) - \hat{m}(0,x)|\mathbb{1}\{\hat{\delta}_+(X) \neq \delta_+(X)\}\right] \\
&\quad + \mathbb{E}\left[(u_g - u_l)|m(1,x) + m(0,x) - 1|\mathbb{1}\{\hat{\delta}_+(X) \neq \delta_+(X)\}\right] \\
&\leq u_g^{\alpha} C(2\|m - \hat{m}\|_{\infty})^{1+\alpha} + u_g C2\|m - \hat{m}\|_{\infty} P(\hat{\delta}_+(X) \neq \delta_+(X)) \\
&\quad + (u_g - u_l) R_+(\hat{\delta}_+).
\end{aligned}
$$

□

References

Athey, S. and Wager, S. (2021). Policy Learning With Observational Data. *Econometrica*, 89(1):133–161.

Audibert, J. Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *Annals of Statistics*, 35(2):608–633.

Babii, A., Chen, X., Ghysels, E., and Kumar, R. (2021). Binary Choice with Asymmetric Loss in a Data-Rich Environment: Theory and an Application to Racial Justice.

Ben-Michael, E., Greiner, D. J., Imai, K., and Jiang, Z. (2021). Safe Policy Learning through Extrapolation: Application to Pre-trial Risk Assessment.

Bertsimas, D., Brown, D. B., and Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501.

Bertsimas, D., Imai, K., and Li, M. L. (2023). Distributionally Robust Causal Inference with Observational Data. *arXiv preprint*, 2.

Bordley, R. F. (2009). The hippocratic oath, effect size, and utility theory. *Medical Decision Making*, 29(3):377–379. PMID: 19380886.

Bradley, R. and Stefánsson, H. O. (2017). Counterfactual desirability. *British Journal for the Philosophy of Science*, 68(2):485–533.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., Fulkerson, W. J., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., and Knaus, W. A. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association*, 276(11):889–897.

Cui, Y. (2021). Individualized decision making under partial identification: three perspectives, two optimality results, and one paradox. *Harvard Data Science Review*. Just accepted.

Cui, Y. and Tchetgen Tchetgen, E. (2021). A Semiparametric Instrumental Variable Approach to Optimal Treatment Regimes Under Endogeneity. *Journal of the American Statistical Association*, 116(533):162–173.

D'Adamo, R. (2023). Orthogonal Policy Learning Under Ambiguity.

Ding, P. and Lu, J. (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 79(3):757–777.

Duchi, J. C. and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, 49(3):1378–1406.

Fan, Y. and Park, S. S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, 5:5–15.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Han, S. (2021). Optimal Dynamic Treatment Regimes and Partial Welfare Ordering.

Heckman, J. J., Smith, J., and Clements, N. (1997). Making the Most out of Programme Evaluations and Social Experiments : Accounting for Heterogeneity in Programme Impacts. *Review of Economic Studies*, 64(4):487–535.

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weight-

ing: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3-4):259–278.

Imai, K. and Jiang, Z. (2023). Principal fairness for human and algorithmic decision-making. *Statistical Science*, 38(2):317–328.

Ishihara, T. and Kitagawa, T. (2021). Evidence Aggregation for Treatment Choice.

Jiang, Z., Ding, P., and Geng, Z. (2016). Principal causal effect identification and surrogate end point evaluation by multiple trials. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(4):829–848.

Jiang, Z., Yang, S., and Ding, P. (2022). Multiply robust estimation of causal effects under principal ignorability. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1423–1445.

Jonsen, A. R. (1978). Do no harm. *Annals of Internal Medicine*, 88(6):827–832.

Kallus, N. (2018). Balanced policy evaluation and learning. *Advances in Neural Information Processing Systems*, 2018-December(1):8895–8906.

Kallus, N. (2022). What's the Harm? Sharp Bounds on the Fraction Negatively Affected by Treatment.

Kallus, N. and Zhou, A. (2021). Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890.

Kennedy, E. H. (2022). Towards optimal doubly robust estimation of heterogeneous causal effects.

Kitagawa, T. and Tetenov, A. (2018). Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*, 86(2):591–616.

Knaus, M. C. (2020). Double Machine Learning based program evaluation under unconfoundedness. *arXiv*, pages 1–61.

Lin, P. (2016). *Autonomous Driving: Technical, Legal and Social Aspects*, chapter Why Ethics Matters for Autonomous Cars, pages 69–85. Springer Open.

Manski, C. F. (2004). Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, 72(4):1221–1246.

Manski, C. F. (2005). *Social Choice with Partial Knowledge of Treatment Response*. Princeton University Press.

Manski, C. F. (2007). Minimax-regret treatment choice with missing outcome data. *Journal of Econometrics*, 139(1):105–115.

Manski, C. F. (2011). Choosing treatment policies under ambiguity. *Annual Review of Economics*, 3:25–49.

Pu, H. and Zhang, B. (2021). Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach. *Journal of the Royal Statistical Society Series B*, pages 1–28.

Qiu, H., Carone, M., Sadikova, E., Petukhova, M., Kessler, R. C., and Luedtke, A. (2021). Optimal Individualized Decision Rules Using Instrumental Variable Methods. *Journal of the American Statistical Association*, 116(533):174–191.

Rüschendorf, L. (1981). Sharpness of Fréchet-bounds. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(2):293–302.

Smith, C. M. (2005). Origin and uses of primum non nocere — above all, do no harm! *Journal of Clinical Pharmacology*, 45:371–377.

Stefánsson, H. O. (2015). Fair Chance and Modal Consequentialism. *Economics and Philosophy*, 31(3):371–395.

Stoye, J. (2012). Minimax regret treatment choice with covariates or with limited validity of experiments. *Journal of Econometrics*, 166(1):138–156.

Sverdrup, E., Kanodia, A., Zhou, Z., Athey, S., and Wager, S. (2020). policytree: Policy learning via doubly robust empirical welfare maximization over trees. *Journal of Open Source Software*, 5(50):2232.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, 59(2):204–217. Philosophical Problems of Death.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., and Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25:1337–1340.

Yata, K. (2021). Optimal Decision Rules Under Partial Identification.

Zhang, Y., Ben-Michael, E., and Imai, K. (2023). Safe Policy Learning under Regression Discontinuity Designs with Multiple Cutoffs.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.